Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion (ISSN 1617-1837)
Ausgabe 15 (2014), Seite 196-233 (www.gespraechsforschung-ozs.de)

Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD

Thomas Schmidt

Abstract
Dieser Beitrag stellt das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) und die Datenbank für Gesprochenes Deutsch (DGD) als Instrumente gesprächsanalytischer Arbeit vor. Nach einer allgemeinen Einführung in FOLK und DGD im zweiten Abschnitt werden im dritten Abschnitt die methodischen Beziehungen zwischen Korpuslinguistik und Gesprächsforschung und die Herausforderungen, die sich bei der Begegnung dieser beiden Herangehensweisen an authentisches Sprachmaterial stellen, kurz skizziert. Der vierte Abschnitt illustriert dann ausgehend vom Beispiel der Formel *ich sag mal*, wie eine korpus- und datenbankgesteuerte Analyse zur Untersuchung von Gesprächsphänomenen beitragen kann.

Keywords: Gesprächskorpus, Mündliches Korpus, Gesprächsdatenbank, Korpuslinguistik, Diskursmarker

English abstract
This paper introduces the Research and Teaching Corpus of Spoken German (FOLK) and the Database for Spoken German (DGD) as tools for doing conversation analysis. After a general introduction to FOLK and the DGD in section 2, section 3 briefly discusses the methodological relationship between corpus linguistics and conversation analysis and the challenges that arise where these two approaches interact. Using formula *ich sag mal* ('let me say') as an example, section 4 then illustrates how a corpus and database driven analysis can contribute to the study of conversation phenomena.

Keywords: Conversation corpus, oral corpus, conversation database, corpus linguistics, discourse marker

1. Einleitung

Dieser Beitrag stellt das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) und die Datenbank gesprochenes Deutsch (DGD)[1] als Instrumente gesprächsanalytischer Arbeit vor. Er tut dies in der Annahme, dass die Gesprächsanalyse, wenngleich sie schon immer grundsätzlich und wesentlich empirisch fundiert war, die Möglichkeiten computergestützter Analysen größerer Datenmengen bislang nur zögerlich erkundet hat – im Gegensatz zur Korpuslinguistik, für die der Einsatz digitaler Technologien von Beginn an charakteristisch war, die sich dafür aber bislang nur am Rande mit authentischen Gesprächsdaten beschäftigt hat.

Nach einer allgemeinen Vorstellung von FOLK und DGD im folgenden Abschnitt werden im dritten Abschnitt die methodischen Beziehungen zwischen Korpuslinguistik und Gesprächsforschung und die Herausforderungen, die sich bei der Begegnung dieser beiden Herangehensweisen an authentisches Sprachmaterial stellen, kurz skizziert. Der vierte Abschnitt illustriert dann ausgehend vom Beispiel der Formel *ich sag mal* und verwandter Konstruktionen, wie eine korpus- und datenbankgesteuerte Analyse zur Untersuchung von Gesprächsphänomenen beitragen kann. Die Darstellung im vierten Abschnitt ist bewusst ausführlich gehalten und behandelt auch technische Fragen der Bedienung der DGD so detailliert, dass sich die Vorgehensweise vom Leser Schritt für Schritt nachvollziehen und gegebenenfalls auf eigene Forschungsfragen übertragen lassen sollte.

Ziel des Beitrags ist in erster Linie das Aufzeigen von Möglichkeiten (und Grenzen) gesprächsanalytischer Korpusanalysen, wie sie derzeit mit FOLK über die DGD umgesetzt werden können. Nicht näher eingehen werde ich hingegen auf einige weitere Bereiche, die mit FOLK und DGD in Zusammenhang stehen und gleichfalls für ein gesprächsanalytisches Publikum von Interesse sein mögen, aber den Rahmen dieses Beitrags sprengen würden. Dazu gehören erstens verschiedene Aspekte des Workflows, der bei der Erstellung von FOLK zum Einsatz kommt und z.B. hierfür spezialisierte Transkriptions- und Annotations-Tools (FOLKER und OrthoNormal, Schmidt/Schütte 2010 und Schmidt 2012), das AGD-Metadatenmodell (Gasch et al. 2008) und Methoden für das automatisierte Lemmati-

[1] http://dgd.ids-mannheim.de

sieren und Part-Of-Speech-Tagging von Gesprächsdaten (Westpfahl/Schmidt 2013) beinhaltet. Zweitens gehe ich auch nicht detaillierter auf alle Aspekte des FOLK-Korpusdesigns und damit zusammenhängende Fragen der Stratifikation, Repräsentativität und Ausgewogenheit von Gesprächskorpora ein (siehe dazu Deppermann/Hartung 2012 und Kupietz/Schmidt 2015). Drittens schließlich blendet dieser Beitrag auch die Rolle, die FOLK, DGD bzw. das Archiv für Gesprochenes Deutsch in Bezug auf die Nachhaltigkeit und Nachnutzbarkeit von Gesprächsdaten insgesamt spielen (siehe Schmidt 2005 oder Stift/Schmidt 2013 für zwei Beiträge, die diesen Aspekt mehr in den Vordergrund rücken), weitestgehend aus.

2. FOLK und DGD2

2.1. FOLK als Gesprächskorpus

In Deppermann/Schmidt (2014:4) haben wir ein Gesprächskorpus definiert als

> […] eine Sammlung von Aufzeichnungen (Audio- und/oder Videoaufnahmen) authentischer Gespräche (i.e. konzeptionell und medial mündlicher, i.d.R. spontaner, Interaktion von zwei oder mehr Teilnehmern), die nach einer wissenschaftlich begründeten und explizit dargelegten Systematik zusammengestellt und über eine Transkription, gegebenenfalls zusätzliche Annotationen und die Dokumentation von Metadaten (zu Gesprächsumständen und beteiligten Sprechern) für eine (sprach-)wissenschaftliche Analyse erschlossen wird.

Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) ist ein Gesprächskorpus, das eine größtmögliche Variation in Bezug auf den Interaktionstyp anstrebt.

In der zentralen Rolle, die das Konzept 'Gespräch' für das Design von FOLK spielt, besteht zum einen ein wesentlicher Unterschied zu vielen anderen Korpora, die zwar mündliche Sprache beinhalten, diese aber nicht (oder zumindest nicht primär) in Form authentischer Gespräche erheben oder als solche im Korpus repräsentieren. Dies ist z.B. überall dort der Fall, wo mündliche Sprache vorwiegend elizitiert wird, etwa in Form biographischer Interviews (vgl. etwa Dialektkorpora wie das Korpus 'Deutsche Mundarten' oder die soziolinguistischen Interviews von William Labov), aber auch dort, wo beliebige Gesprächsausschnitte lediglich als Instanzen von 'Spontansprache' Eingang in Datensammlungen finden, die Tatsache, dass sie Bestandteile von Gesprächen sind (und damit auch die Dokumentation des Gesprächskontexts), also in den Hintergrund tritt.

Zum anderen unterscheidet sich FOLK in seinem Bemühen um eine große Variationsbreite auch von Gesprächskorpora,[2] die sich auf spezifische Interaktionstypen (z.B. Interaktionen an der Universität in GeWiss, Fandrych et al. 2012 oder MICASE, Simpson-Vlach/Leicher 2006, Arzt-Patienten-Gespräche im Krankenhaus in DiK, Bührig et al. 2012) und/oder spezifische Teilnehmer oder Teilnehmerkonstellationen (z.B. multiethnische Sprechergemeinschaften in Berlin im

[2] Ich beschränke mich hier und im Folgenden auf Korpora, die im Idealfall veröffentlicht sind, zu denen aber mindestens ausreichend publizierte Information vorliegt. Dass damit viele existierende gesprächsanalytische Datensammlungen ausgeschlossen bleiben, ist mir bewusst.

KidKo-Korpus, Wiese et al. 2012, rezeptiv Mehrsprachige in Skandinavien in SkandSemiko, Schmidt 2003) fokussieren.

Die "(sprach-)wissenschaftliche Analyse", von der in der Definition die Rede ist, kann selbstverständlich eine gesprächsanalytische sein, und FOLK ist in der Tat primär an den Bedürfnissen von Gesprächsforschern ausgerichtet. Als öffentlich verfügbares Korpus adressiert FOLK zum einen den auch in der Gesprächsforschung verschiedentlich beklagten (siehe etwa Schmidt 2005) Mangel an adäquat aufbereiteten, rechtlich autorisierten Daten, die von einem größeren Nutzerkreis verwendet werden können. Zum anderen bietet es als stetig wachsendes und systematisch stratifiziertes Korpus auch zum ersten Mal einen Referenzpunkt, von dem aus sich gesprächsanalytische Einzelfallanalysen über einen Rückgriff auf eine größere Datenbasis bewerten und absichern lassen.

2.2. Auf- und Ausbau von FOLK

Das Projekt FOLK wurde 2008 an der Abteilung Pragmatik des IDS Mannheim ins Leben gerufen. In der ersten Erhebungsphase wurden Kontakte von Projektmitarbeitern sowie von Teilnehmern einer regelmäßig abgehaltenen Lehrveranstaltung genutzt, um geeignete Gesprächsanlässe zu identifizieren und aufzeichnen zu lassen. Ergänzt wurde dies durch Aufnahmen aus dem Korpus 'Deutsch heute' des IDS-Projekts 'Variation des gesprochenen Deutsch' (Maptasks und sprachbiografische Interviews, Brinckmann et al. 2008) sowie durch Datenspenden externer Projekte (insbesondere universitäre Prüfungsgespräche aus dem GeWiss-Projekt, Fandrych et al 2012). Um die Variationsbreite zu erhöhen und die regionale Verteilung der Aufnahmen auszugleichen, werden in der derzeitigen Ausbauphase verstärkt Kooperationen mit Institutionen in anderen Teilen der Bundesrepublik gesucht, die Expertise in der Erhebung mündlicher Daten haben und persönliche Kontakte vor Ort für einen geeigneten Feldzugang nutzen. Daten aus einer solchen Kooperation mit dem Hamburger Zentrum für Sprachkorpora[3] und dem Projekt 'Gesprochene Sprache im Ruhrgebiet'[4] werden derzeit aufbereitet, eine weitere Kooperation mit dem Arbeitsbereich Korpuslinguistik der HU Berlin[5] wurde begonnen. Hinzu kommen weitere Datenspenden, etwa aus den Projekten 'Sprachvariation in Norddeutschland'[6] und 'Theater im Gespräch. Sprachliche Kunstaneignungspraktiken in der Theaterpause'[7] sowie aus den Dissertationsprojekten von Hee (2012) und Weber (2014).

Die Daten aus dem Feld (Aufnahmen, Einverständniserklärungen und Metadaten) werden im Projekt auf Vollständigkeit und rechtliche Autorisierung geprüft, bevor sie in verschiedenen Arbeitsschritten (Schneiden, Bearbeiten und Maskieren der Aufnahmen, elektronische Eingabe der Metadaten, Transkription in FOLKER nach cGAT, Normalisierung, Lemmatisierung und POS-Tagging der Transkription) für das Korpus erschlossen und schließlich über die Datenbank für Gesprochenes Deutsch (DGD, s.u.) veröffentlicht werden.

---

3 https://corpora.uni-hamburg.de/
4 http://www.ruhr-uni-bochum.de/kgsr/
5 https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/
6 http://www.corpora.uni-hamburg.de/sin/startseite.html
7 http://www.uni-siegen.de/phil/lissie/theater_im_gespraech/

Das erste offizielle Release des Korpus erfolgte mit der ersten Version der DGD im Dezember 2012. Mittlerweile (Release vom März 2014) ist die veröffentlichte Version von FOLK auf 101h Audio-Material angewachsen, was knapp einer Million transkribierter Wörter entspricht. Diese setzen sich wie folgt zusammen:

Interaktionstyp	Ereignisse	Tokens	Dauer
Gespräch auf der Urlaubsreise	2	5477	00:28:46
Gespräch beim Umräumen	1	5228	00:21:05
Gespräch in der Familie	2	23414	01:50:50
Gespräch unter Freunden	1	24744	02:17:27
Paargespräch	3	20980	02:41:45
Spielinteraktion mit Kindern	4	40514	05:08:35
Spielinteraktion zwischen Erwachsenen	2	64968	06:41:39
Studentisches Alltagsgespräch	4	42295	03:07:27
Tischgespräch	7	89281	08:01:09
Vorlesen für Kinder	6	18901	02:58:38
Maptask	25	64257	07:15:47
Feedbackgespräch unter Lehrkräften	1	5991	00:24:12
Gespräch im Polizeirevier	9	27515	03:12:11
Meeting in einer sozialen Einrichtung	3	85256	07:34:14
Prüfungsgespräch in der Hochschule	19	98592	10:21:21
Schichtübergabe in einem Krankenhaus	8	28108	02:38:10
Training in einer Hilfsorganisation	9	15217	01:36:02
Unterrichtsstunde im Wirtschaftsgymnasium	8	51760	07:00:13
Unterrichtsstunde in der Berufsschule	7	50050	07:31:26
Schlichtungsgespräch	2	102535	10:28:59
Sprachbiografisches Interview	14	100569	09:33:27
Gesamt	**137**	**965652**	**101:12**

Tabelle 1: Zusammensetzung von FOLK nach Gesprächstypen

In den kommenden Jahren soll FOLK um mindestens 30h Material jährlich anwachsen. Das nächste Release ist für März 2015 geplant.

2.3. Die DGD als Gesprächsdatenbank

Es besteht in der korpuslinguistischen Literatur Einigkeit darüber, dass Möglichkeiten und Grenzen korpusbasierter Untersuchungen nicht nur von den Daten selbst, sondern ganz wesentlich auch von den Instrumenten abhängen, die zu deren Verarbeitung und Analyse zur Verfügung stehen (beide Zitate nach Anthony 2013:144):

> [...] a corpus by itself can do nothing at all, being nothing more than a store of used language (Hunston 2002:20).

> The essence of the corpus as against the text is that you do not observe it directly; instead you use tools of indirect observation, like query languages, concordancers, collocators, parsers, and aligners [...] (Sinclair 2004:189).

Als Mindestanforderung für die Arbeit mit einem Gesprächskorpus mag die Bereitstellung von Audio-Aufnahmen und Transkripten in einer für den Forscher nutzbaren Form – etwa als Download und zum Abspielen bzw. Lesen in einem Audioeditor bzw. einer Textverarbeitung – angesetzt werden (siehe dazu auch

Merkel/Schmidt 2009). Bei größeren Datenmengen offenbaren sich jedoch schnell die Grenzen einer solch einfachen 'Gesprächsdatenbank': abgesehen davon, dass schon der reine Umfang der digitalen Dateien (z.B. knapp 70GB für die aktuelle Version von FOLK) in der Praxis schwer zu handhaben ist, sind ohne weitere gezielte Hilfen zum Erschließen der Daten kaum Analysen möglich, die das Korpus systematisch in seiner Gesamtheit (d.h. nicht nur in mehr oder weniger zufällig gewählten Ausschnitten) einbeziehen. Neben der Bereitstellung der Ausgangsdaten leistet eine zeitgemäße Gesprächsdatenbank daher mindestens folgendes:

- Sie ermöglicht ein Online-Browsing, d.h. ein exploratives Anhören und Lesen der Daten ohne die Notwendigkeit eines Downloads. Dabei sollten auch die Beziehungen zwischen verschiedenen Datentypen (Alignierung zwischen Transkript zu Audio, Metadaten zu Gesprächen und Sprechern, etc.) möglichst einfach abrufbar sein.
- Sie erlaubt es, anhand von Metadaten zu Gesprächen und Sprechern Teilmengen an Datensätzen auszuwählen, die gewissen Kriterien genügen (z.B. Alltagsgespräche mit männlichen Teilnehmern zwischen 20 und 40 Jahren).
- Sie erlaubt ein gezieltes Durchsuchen der Transkripte nach sprachlichen Oberflächenformen (z.B. alle Vorkommen des Verbs *meinen* im Korpus) und stellt die Fundstellen in ihrem Gesprächskontext, gegebenenfalls auch mit zugehörigen Metadaten, dar.

Eine Korpusdatenbank wird umso mehr zu einer Gesprächsdatenbank, je mehr sie gesprächsanalytische Arbeitsweisen ermöglicht. Es gibt Korpusdatenbanken mit Gesprächsdaten, bei denen dies nur sehr eingeschränkt der Fall ist. So wurde etwa bei der Bereitstellung der Korpora 'Grundstrukturen (Freiburger Korpus)' und 'Dialogstrukturen' in früheren Versionen des COSMAS-System des IDS zwar ein gezieltes Durchsuchen der Transkripte, aber weder der Zugriff auf das zugehörige Audio noch auf den Transkriptkontext ermöglicht. Ähnlich verhält es sich bei vielen weiteren Korpusdatenbanken, die mündliche Sprache lediglich in transkribierter Form und damit quasi als spezielle Texte neben weiteren genuin schriftsprachlichen Daten zur Verfügung stellen.

Die Datenbank für Gesprochenes Deutsch (DGD) ist – wie auch ihr Vorgänger (Fiehler/Wagener 2005) – hingegen grundsätzlich für die Arbeit mit mündlichen Daten ausgelegt und bedient daher die oben genannten Anforderungen vollständig. Eine detaillierte Darstellung, wie mit der DGD eine gesprächsanalytische Fragestellung bearbeitet werden kann, findet sich im vierten Abschnitt dieses Beitrags. An dieser Stelle sei daher nur ein kurzer Überblick über die relevante Grundfunktionalität gegeben.

Über den Menüpunkt **KORPORA** ermöglicht die DGD das Ansehen der Metadaten zu Gesprächen und Sprechern, das Anhören von beliebigen Stellen der Aufnahmen, den Download von zum Korpus gehörigen Zusatzmaterialien (wie Wortlisten, Sitzpläne oder Verlaufsprotokolle zu einzelnen Gesprächen, etc.) sowie das Lesen von Transkripten. Insbesondere bei der Darstellung von Transkripten wird Wert darauf gelegt, Daten verschiedenen Typs in eine einzige, leicht zu handhabende Repräsentationsform zu integrieren. So kann durch einen Doppelklick auf eine beliebige Stelle im Transkript der zugehörige Ausschnitt aus der Audioaufnahme abgerufen werden. Eine gestrichelte Linie (1 in Abb. 1) zeigt dann die ak-

tuell abgespielte Position an. Ein Klick auf die Kennung des Transkripts (2 in Abb. 1) zeigt die zur Interaktion dokumentierten Metadaten (Abb. 2 links) an, ein Klick auf ein Sprecherkürzel des Transkripts (3 in Abb. 1) die zugehörigen Sprechermetadaten (Abb. 2 rechts). Außer einer an GAT orientierten Präsentation des Transkripts in Zeilennotation (Abb. 1) steht für FOLK-Daten auch eine Darstellung in Partiturnotation (Abb. 3) zur Verfügung, anhand derer sich zeitliche Verhältnisse, insbesondere Simultaneität, gegebenenfalls besser ablesen lassen.

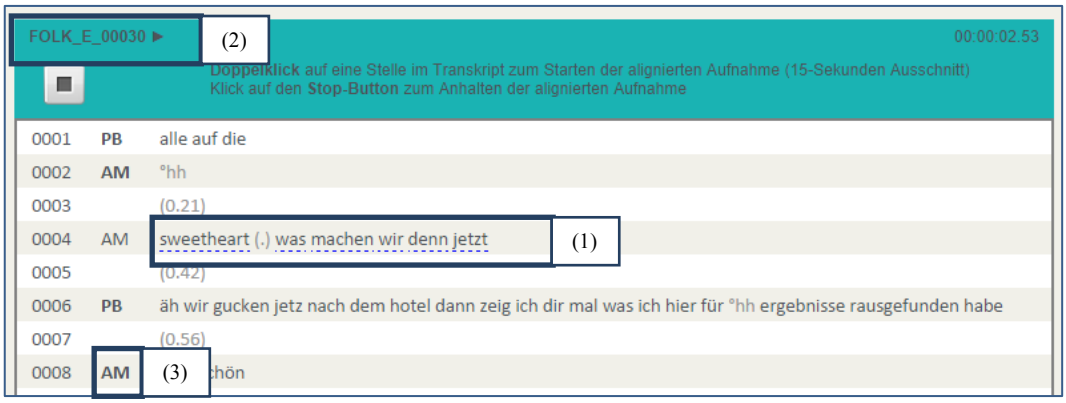

Abbildung 1: Anzeige eines Transkripts aus FOLK in der DGD2[8]

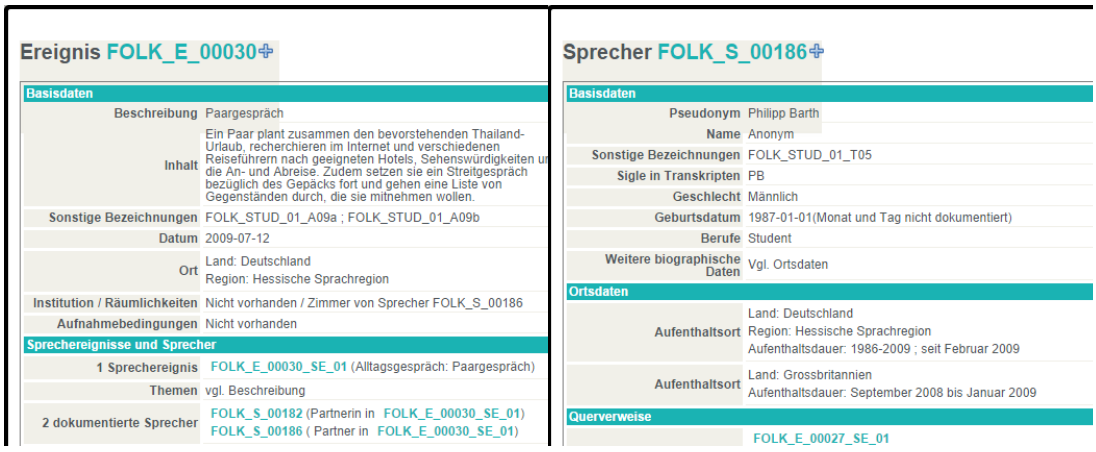

Abbildung 2: Anzeige von Metadaten zum Gespräch (links)[9] und zu einem Sprecher (rechts)[10]

---

[8] Transkript FOLK_E_00030_T_01, auch direkt einsehbar über folgende URL (DGD-Zugangsdaten erforderlich): http://dgd.ids-mannheim.de/service/DGD2Web/ExternalAccessServlet?command=displayData&id=FOLK_E_00030_SE_01_T_01
Im Folgenden wird allen Abbildungen und Beispielen wo immer möglich ein direkter Link auf die DGD beigefügt.

[9] http://dgd.ids-mannheim.de/service/DGD2Web/ExternalAccessServlet?command=displayData&id=FOLK_E_00030

[10] http://dgd.ids-mannheim.de/service/DGD2Web/ExternalAccessServlet?command=displayData&id=FOLK_S_00186

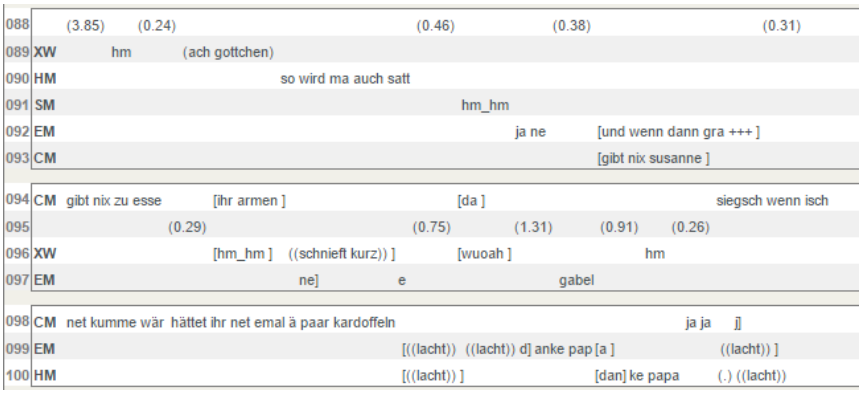

Abbildung 3: Anzeige eines Transkripts in Partiturnotation

Über den Menüpunkt **RECHERCHE > METADATEN** lässt sich der Datenbestand nach den verschiedenen Informationen durchsuchen, die in den Metadaten zu Gesprächen und Sprechern dokumentiert sind. In Abb. 4 und 5 wurde auf diese Weise ein Teilkorpus gebildet, das nur Gespräche aus dem privaten Bereich enthält, an denen mindestens ein männlicher Sprecher im Alter zwischen 20 und 40 Jahren beteiligt ist. Ein solches Teilkorpus kann als sogenanntes virtuelles Korpus dauerhaft in der Datenbank gespeichert und z.B. als Grundlage einer wieteren Suche auf den zugehörigen Transkripten verwendet werden.

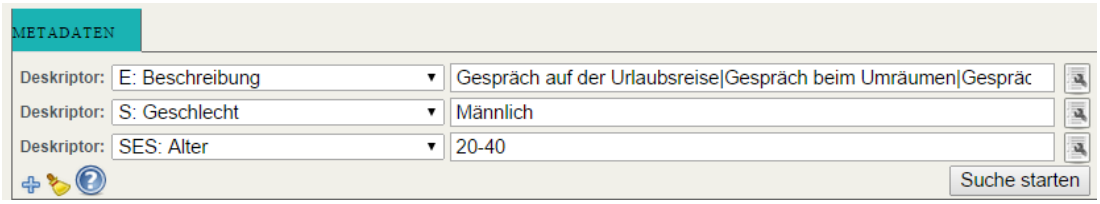

Abbildung 4: Suchkriterien für die Zusammenstellung eines virtuellen Korpus

Abbildung 5: Virtuelles Korpus als Ergebnis der Suche aus Abb. 4

Über den Menüpunkt **RECHERCHE > TOKENS** können die Transkripte eines Korpus oder Teilkorpus nach bestimmten Formen durchsucht werden. Als Suchbegriffe können transkribierte Formen (in literarischer Umschrift), normalisierte Formen (in Standardorthographie) oder Lemmata (i.e. Grundformen) – oder eine Kombination dieser Formen – verwendet werden. In Abbildung 6 wurde z.B. eine Suchanfrage nach dem Lemma *meinen* gestellt, die als Ergebnis auch alle

flektierten Formen des Verbs (*meinte*, *gemeint*, *meintest*, …), darunter auch solche Formen, deren Transkription von der Standardorthographie abweicht (*mein*, *gmeint*), liefert – nicht aber Formen des Possessivpronomens *mein* (*meinen*, *meiner*, …). Die Ergebnisse werden als KWIC (*Keyword in Context Concordance*, Abb. 7) angezeigt. Dabei können wiederum die Stelle der zugehörigen Aufnahme abgespielt, die Fundstelle im Transkriptkontext angezeigt und Metadaten zu Gespräch und Sprechern direkt aus der Konkordanz heraus abgerufen werden. Auf diese Weise wird ein manuell-intellektuelles Inspizieren einzelner Fundstellen unterstützt, was z.B. ein Aussortieren von *false positives* (s.u.) oder eine Bewertung von Fundstellen anhand prosodischer Informationen (die im Transkript nicht enthalten sind und insofern einen Rückgriff auf die Aufnahme erfordern) ermöglicht.

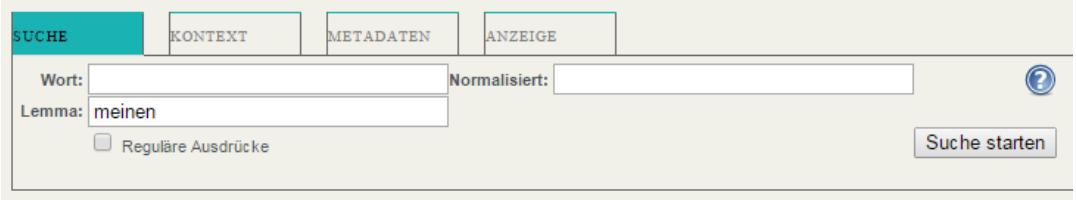

Abbildung 6: Suchkriterien für die Token-Suche nach dem Lemma meinen

Abbildung 7: Ergebnis der Token-Suche als KWIC-Konkordanz mit expandiertem Transkriptausschnitt

3. Daten und Methoden zwischen Korpuslinguistik und Gesprächsforschung

Methodisch bedienen Gesprächsdatenbanken eine Schnittstelle zwischen Korpuslinguistik und Gesprächsforschung. Beide Herangehensweisen sind in einem fundamentalen Sinne empirische Ansätze zur Analyse sprachlicher Phänomene, d.h. sie nehmen authentische Daten als Ausgangspunkt ihrer Untersuchung und legen Wert darauf, die Analyse von diesen Daten und nicht von à priori gebildeten Kategorien leiten zu lassen. Dennoch unterscheiden sich die beiden Bereiche sowohl in Hinblick auf ihr Datenverständnis als auch hinsichtlich ihrer methodischen Grundsätze erheblich.

Betrachtet man erstens das Datenverständnis in der Korpuslinguistik, so fällt auf, dass das Verhältnis zwischen Primärdaten und deren Repräsentation im Korpus kaum thematisiert wird. Vielmehr werden das (i.d.R. schriftsprachliche) Aus-

gangsmaterial und dessen (i.d.R. elektronische) Repräsentation im Korpus – wie im folgenden Zitat aus Lemnitzer/Zinsmeister (2006:7) unter dem Terminus 'Text', der dann auch ohne weitere Unterscheidung gesprochene Sprache subsumiert, weitestgehend gleichgesetzt:

> Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind.

Demgegenüber steht in der Gesprächsforschung eine tiefgehende Auseinandersetzung mit dem theorie- oder modellhaften Status des Transkripts. Es scheint mir an dieser Stelle nicht notwendig, dies weiter zu erläutern – ein Verweis auf die grundlegende Arbeit von Ochs (1979) oder verschiedene Einführungen in die Gesprächsanalyse (z.B. Deppermann 1999) mögen genügen, um daran zu erinnern, dass die Gesprächsforschung weit davon entfernt ist, das empirische Ausgangsmaterial (Aufnahmen authentischer Gespräche) und dessen textförmige Repräsentation in einem Korpus gleichzusetzen.[11]

Zu erklären ist das unterschiedliche Gewicht, das Korpuslinguistik und Gesprächsforschung auf diese Thematik legen, mit den Daten und Erkenntnisinteressen selbst. Zwar fließen auch in die Digitalisierung schriftsprachlicher Texte Abstraktionen und Interpretationen ein (etwa wenn Layout und Formatierung eines Zeitungstextes bei dessen Überführung in eine Textdatei "verloren gehen"), diese haben aber nur geringe Auswirkungen auf die Analysetätigkeit, da sie die sprachliche "Essenz" des Ausgangsmaterials (d.h. die lineare Abfolge diskreter Wörter und Interpunktionszeichen) weitestgehend vollständig und unverändert lassen. Dies ist bei der Überführung von mündlicher Sprache in schriftliche Transkripte grundsätzlich anders, da dort zwangsläufig Entscheidungen getroffen werden müssen, wie Kontinuierliches (wie etwa prosodische Verläufe) auf Diskretes (i.e. Transkriptionszeichen) und Simultanes (wie Überlappungen) auf Lineares (i.e. den Transkript-"Text") abgebildet werden soll. Anders als in der Korpuslinguistik der Standard-Schriftsprache kann daher bei einer korpuslinguistischen Untersuchung von Gesprächsdaten das Korpus nicht einfach als eine vollständige und eindeutige Repräsentation des Untersuchungsgegenstandes behandelt werden; es muss immer in seiner Relation zur Aufnahme und vor dem Hintergrund der Entscheidungen, die Transkriptionssystem bzw. Transkribent bei der Verschriftlichung der Aufnahmen treffen, gesehen werden.

Gesprächsanalytische Transkripte verhalten sich zweitens auch in Bezug auf ihre automatische Durchsuchbarkeit anders als digitalisierte standardsprachliche Texte. Bei letzteren sorgt mit der Standard-Orthographie eine externe Norm dafür, dass Nutzer eine klare und eindeutige Erwartung an die im Korpus zu suchenden Formen haben und dass folglich ein versehentliches Übergehen von relevanten Belegen (*false negatives*) in der Regel ausgeschlossen ist. Transkripte hingegen stellen immer einen Kompromiss zwischen Vereinheitlichung (u.a. in Hinblick auf eine bessere "Lesbarkeit") und detaillierter Wiedergabe von Variation dar.

[11] In einer recht radikalen Auslegung beschreibt (ten Have 1990:2) das Verhältnis zwischen Aufnahme und Transkript so: "Recordings are CA's basic data. The transcriptions made after these are to be seen as a convenient form to represent the recorded material in written form, but not as a real substitute".

Augenfällig ist dies besonders beim Einsatz der literarischen Umschrift, die einerseits verwendet wird, um Variation in der lautlichen Realisierung von Wörtern wiederzugeben, sich andererseits aber an die Standardorthographie anlehnt, um schriftsprachlichen Lesegewohnheiten dennoch soweit wie möglich entgegenzukommen. Bei einem Korpus wie FOLK, das Daten aus unterschiedlichsten Registern und Regionen enthält und von einer Vielzahl von studentischen Hilfskräften transkribiert wird, ist die Heterogenität, die über die literarische Umschrift eingeführt wird, trotz aller Bemühungen um Vereinheitlichung sehr groß. Es lassen sich daher kaum mehr verlässliche Erwartungen für eine Suche formulieren. Beispielsweise finden sich für das Lexem *nein* nicht weniger als 9 verschiedene literarisch transkribierte Formen (*nein*, *nee*, *na*, *ne*, *neeh*, *nehee*, *nö*, *näh* und *nää*) in der aktuellen Version des Korpus, von denen zumindest ein Teil nicht ohne Weiteres vorhersagbar ist.

Schließlich besteht drittens ein wesentlicher Unterschied zwischen den beiden Herangehensweisen in ihrer Haltung zum Quantifizieren. Bei der Verwendung großer Datenmengen, die entsprechend große Belegzahlen für ein zu untersuchendes Phänomen liefern, drängt sich die Frage mehr oder weniger von selbst auf, ob diese Zahlen sich nicht für statistisch begründete Generalisierungen nutzen lassen. Die Korpuslinguistik steht daher, auch wenn sie keinesfalls eine rein quantitative Methode ist, quantitativ begründeten Befunden offen gegenüber, solange diese durch entsprechende Überlegungen zu Repräsentativität und Ausgewogenheit der Datenbasis abgesichert sind (vgl. z.B. Biber 2004). Die Gesprächsforschung hingegen ist auf die detaillierte, rein qualitative Analyse von Einzelfällen orientiert und kann daher als generell "quantifizierungsskeptisch" bezeichnet werden. Hinzu kommt, dass es ihr aufgrund der spärlichen Datenlage meist auch an praktischen Möglichkeiten fehlt, quantitativ aussagekräftige Analyseergebnisse zu erhalten.

Wenn sich also korpuslinguistische und gesprächsanalytische Methoden in einer Gesprächsdatenbank begegnen und sinnvoll kombinieren lassen sollen, reicht es weder aus, gesprächsanalytische Daten einfach in durchsuchbare Datensätze zu überführen, noch können Techniken, die für schriftsprachliche Korpusanalysen entwickelt wurden, einfach unverändert auf mündliche Daten übertragen werden. In FOLK und DGD werden folgende Maßnahmen getroffen, um eine bessere Vereinbarkeit der beiden Herangehensweisen zu erreichen:

- Wie bereits in Abschnitt 2.3 erwähnt, bietet die DGD wo immer möglich einen direkten Rückgriff vom Transkript oder von einem Suchergebnis in die zugehörige Stelle der Aufnahme. Dies stellt sicher, dass Nutzer sich jederzeit selbst der Ädaquatheit der Transkription versichern können. Darüber hinaus erlaubt die DGD den Download beliebiger Transkriptausschnitte mit dem zugehörigen Audio von bis zu einer Minute Länge. So kann ein Datensatz bei Bedarf auf dem eigenen Rechner weiterbearbeitet werden, etwa indem das GAT-Minimaltranskript zu einem Basistranskript ausgebaut oder eine Software wie Praat zu akustischen Messungen eingesetzt wird. Beides trägt dem Status des Transkripts als zweckgerichtet reduziertem Modell der Ausgangsdaten Rechnung und versetzt den Nutzer in die Lage, dieses Modell gegebenenfalls auf seine eigenen Analysezwecke hin anzupassen.
- Zusätzlich zur literarischen Umschrift werden in FOLK drei weitere Annotationsebenen hinzugefügt. Darunter nimmt die orthographische Normalisie-

rung, also die Abbildung literarisch transkribierter Formen auf ihre standardorthographischen Entsprechungen, eine zentrale Rolle ein. Sie stellt einerseits sicher, dass Suchanfragen gemäß den an der Schriftsprache gebildeten Erwartungen gestellt werden können. Es genügt z.B. eine Suche nach der normalisierten Form *nein*, um alle oben aufgeführten Varianten dieses Lexems aufzufinden. Andererseits schafft sie die Voraussetzung, um – ebenfalls anhand der Schriftsprache entwickelte – automatische Annotationsmethoden (z.B. Tagger) für die Lemmatisierung und Part-Of-Speech-Annotation der Daten einsetzen zu können. In diesem Sinne stellt die Normalisierung also eine Brücke dar, die gesprächsanalytische Daten mit korpuslinguistischen Tools verbindet.

Schließlich sei noch darauf hingewiesen, dass die DGD dem Gesprächsforscher auch die Möglichkeit lässt, FOLK-Daten zu nutzen, ohne sich näher auf korpuslinguistische Methoden einzulassen. Eine "klassische" Transkriptstudie ist auch im Browsing-Modus der DGD ohne weiteres möglich und dort eventuell sogar komfortabler als mit anderen Hilfsmitteln durchzuführen, da auch hier jederzeit die Verknüpfung zwischen Transkript und Audio genutzt werden kann. Ein rein manuelles Zusammenstellen von Transkriptausschnitten (d.h. beim Lesen von Transkripten und ohne vorausgehende automatische Suche) wird außerdem durch die sogenannte Kollektionen-Funktionalität (siehe Abb. 8) unterstützt. Diese erlaubt dem Nutzer, durch einen Klick auf eine Zeilennummer im Transkript den betreffenden Ausschnitt einer Liste hinzuzufügen, dauerhaft in der Datenbank zu speichern und dann jederzeit (über MEINE DGD > KOLLEKTIONEN) abzurufen.

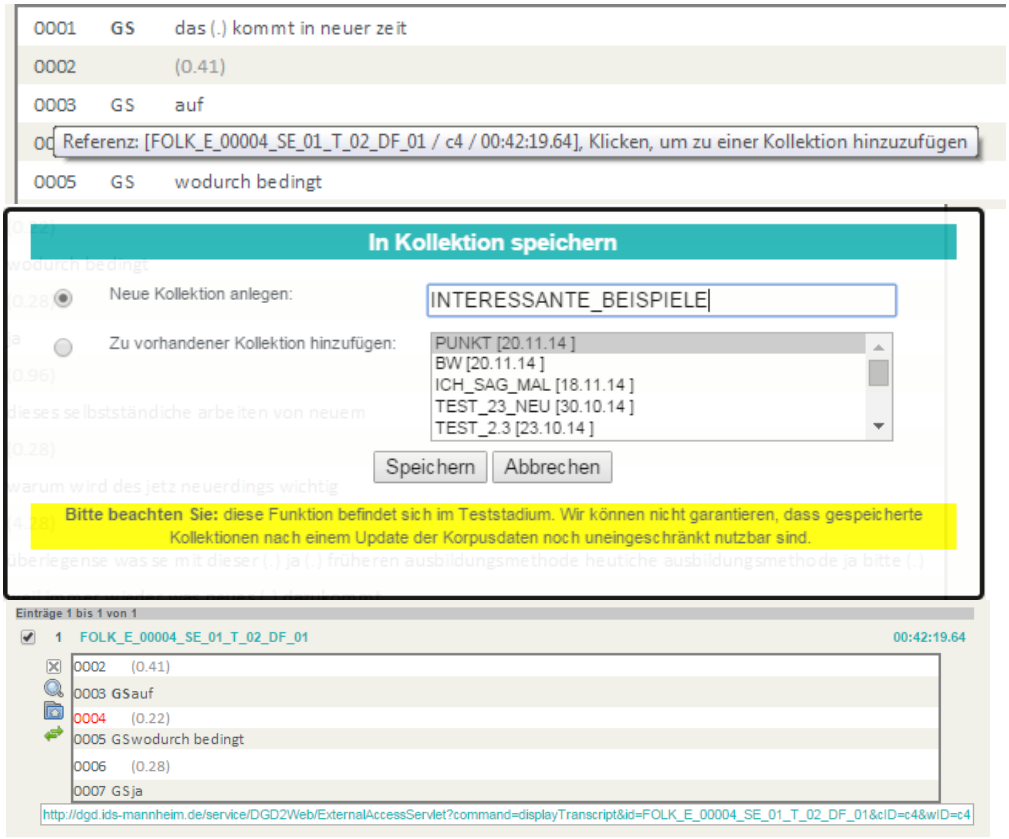

Abbildung 8: Speichern und Abrufen von Transkriptausschnitten in Kollektionen

4. Beispielanalyse *mal sagen*

4.1. *ich sag mal* als Diskursmarker

Als Ausgangspunkt, um die Möglichkeiten von Gesprächskorpora und Gesprächsdatenbanken näher zu illustrieren, soll der folgende Analyseausschnitt aus Auer/ Günthner (2003:12) dienen:

> Ähnlich wie *weiß ich nicht* wird gelegentlich auch *ichsachma(so)* (> *ich sag mal (so)*) als Diskursmarker zur Verzögerung verwendet. Im folgenden Ausschnitt […] leitet der Sprecher mit *ick=sag=mal=so* bzw. *ick=sach=mal=so* (Z. 56, 59 und 62) stark evaluierende Äußerungen ein. Der Diskursmarker *ick=sag=mal=so* fungiert hier als eine Abschwächung der kommenden Äußerung bzw. Evaluation:

```
10) BRANDENBURG 9

55 Udo: ach das is iDIOtisch;
56      ick- ick- ick=sag=mal=so,
57      will- willst du denn all- all- allet solche
        iDIOten
58      am bundestach äh sitzen haben;
59      ick=sach=mal=so,
60      die mußte doch jut bezahlen;
61      die leute die- die muß (..) jut bezahlen;
62      ick=sag=mal=so,
63      son- sonst kannste da irgendwelche idioten
        einsetzen.
```

> In beiden Fällen kann die Verzögerung selbst strategisch sein und - wie in Bsp. (9) und (10) - kritische (im Fall von Bsp. 9 politisch nicht korrekte) Äußerungen einleiten.

Die Konstruktion *ich sag mal (so)* wird hier als ein "Matrixsatz mit *Verbum dicendi*" analysiert, der einen "Übergang zum Diskursmarker" vollzogen hat, wobei als eine hierfür wesentliche Eigenschaft "die Tatsache, dass diese und andere Verben unter bestimmten Bedingungen […] keinen nachfolgenden eingebetteten dass-Satz erfordern, sondern von einem Syntagma mit Hauptsatzsyntax gefolgt werden können", genannt wird (ebda.:9). Weiterhin wird als Charakteristikum solcher Konstruktionen angeführt, dass "der semantische Gehalt des Verbs […] in dieser Verwendungsweise verblasst [ist]" (ebda.:10) bzw. "[d]ie Bedeutung der einzelnen aus Matrixsätzen reduzierten Lexeme […] nicht ohne weiteres aus der Semantik des jeweiligen Vollverbs ableitbar [ist]" (ebda.:11). Für Diskursmarker des Deutschen im Allgemeinen wird außerdem festgestellt, dass "[diese] topologisch durch ihre 'periphere' syntaktische Stellung gekennzeichnet (die sie u.a. von den Modalpartikeln unterscheidet) [sind]: sie sind selbständigen Syntagmen voran- oder nachgestellt" (ebda.:1).

Diese Analyse soll hier nicht in Frage gestellt werden, und es soll auch nicht bestritten werden, dass das angeführte authentische Beispiel die Analyse treffend illustriert. Dennoch lässt die Beschränkung auf ein einzelnes Beispiel interessante Fragen offen, die sich über eine Konsultation größerer Datenmengen möglicherweise beantworten lassen:

- Was ist unter einer "gelegentlichen" Verwendung dieser Konstruktion zu verstehen? Ist sie allgemein selten oder seltener als verwandte Konstruktionen? Ist sie eventuell auf bestimmte Gesprächs- oder Sequenztypen, bestimmte Sprecher oder bestimmte regionale Varianten des Deutschen beschränkt?
- Wie fixiert oder variabel ist die Form dieser Konstruktion? Ist sie – wie die Transkriptionsweise mit den Verschleifungszeichen im Beispiel vielleicht nahelegt – soweit als Diskursmarker "grammatikalisiert", dass keine morphologischen und/oder syntaktischen Varianten mit derselben Funktion vorkommen? Kommt die Konstruktion auch in anderen syntaktischen Positionen vor?
- Wie typisch ist die durch das Beispiel illustrierte Verwendung von *mal sagen*? Inwieweit konkurriert sie mit Verwendungen von *ich sag mal*, die eher als "echte" Matrixsätze mit *verbum dicendi* zu analysieren wären? Sind alle diskursmarkerartigen Verwendungen von *mal sagen* als "Abschwächung" treffend beschrieben oder gibt es verwandte Formen, die andere meta-pragmatische Funktionen erfüllen? Sind andere Verwendungen von *ich sag mal* immer eindeutig abzugrenzen von der im Beispiel illustrierten Verwendung?

In einem Seminar wurden Studierende vor einer Korpusanalyse zu ihren diese Fragen betreffenden Intuitionen befragt. Einige interessante Antworten lauteten wie folgt:

- neben *ich sag mal* existiert auch die Form *sag ich mal* mit gleicher Funktion
- die Form tritt gehäuft in Interviews mit Profifußballern / in Talkshows auf
- die Form ist "typisch süd(west)deutsch"
- außer *so* werden auch *einfach* und *jetzt* häufig als Ergänzungen der Konstruktion verwendet, teilweise auch in Kombination (*ich sag jetzt einfach mal so*)
- verwandte Formen sind *ich würde mal sagen* und *sagen wir mal*, wobei nicht sicher ist, ob sie die gleiche Funktion erfüllen
- nicht reduzierte Formen (*ich sage einmal*), Formen mit anderem Tempus (*ich sagte mal*) oder Formen in der zweiten oder dritten Person (*du sagst mal*) sind in dieser Funktion schwer vorstellbar

Im Folgenden soll diesen Fragen und Intuitionen anhand einer Korpusanalyse von FOLK in der DGD2 nachgegangen werden.

4.2. Analyse von *mal sagen* in FOLK

4.2.1. Initiale Suchanfrage

Als ersten Schritt benötigen wir einen Suchausdruck, der uns möglichst alle für die Fragestellung relevanten Formen (also keine *false negatives*), aber möglichst wenige Formen darüber hinaus (*false positives*) liefert. Wenn wir die Kookkurrenz von *(ein)mal* und (Formen von) *sagen* als das invariante Charakteristikum der uns interessierenden Konstruktionen definieren (i.e. Formen wie *ich sag* oder *ich denk mal* zumindest vorerst unberücksichtigt lassen), läuft dies darauf hinaus, dass wir ein geeignetes Kookkurrenzfenster bestimmen müssen, innerhalb dessen

(fast) alle relevanten Konstruktionen zu erwarten sind. Wir beginnen daher wie folgt:

Um eine Token-Suche nach dem Lemma *sagen* in FOLK auszuführen, rufen wir den Menüpunkt **RECHERCHE > TOKENS** auf, wählen dort FOLK als Korpus in der Liste am linken Bildschirmrand aus und geben die Form *sagen* im Feld "Lemma" des Reiters **SUCHE** ein. Ein Klick auf den Button **SUCHE STARTEN** liefert 5679 Treffer, die als KWIC angezeigt werden.

Abbildung 9: Token-Suche nach dem Lemma ″sagen″

Um nun den Kontext nach den Formen *mal* oder *einmal* zu filtern, wechseln wir in den Reiter **KONTEXT**. Analog zur initialen Suche können hier Eigenschaften von im Kontext der vorhandenen Treffer zu suchenden Token spezifiziert werden. Da wir uns für zwei unterschiedliche Formen interessieren, aktivieren wir die Option "Reguläre Ausdrücke" und geben "(ein)?mal" (zu lesen als: *mal* mit optional vorangestelltem *ein*) im Feld "Normalisiert"[12] ein. Zusätzlich spezifizieren wir Parameter für das zu verwendende Kontextfenster. Wir wählen dieses zunächst bewusst großzügig, um zu ermitteln, welches der maximale Abstand von *mal* und *sagen* bei den uns interessierenden Formen ist. Da wir einstweilen nicht ausschließen wollen, dass auch Varianten mit vorangestelltem *mal* existieren (*mal sag ich*?), suchen wir also *mal* im rechten und linken Kontext von *sagen* mit einem maximalen Abstand von 15 Tokens. Wir beschränken uns auf die Suche innerhalb von Sprecherbeiträgen ("Skopus: Beitrag"), da wir nicht erwarten, dass kookkurrierende *mal* und *sagen* sich in relevanten Belegen auf unterschiedliche Beiträge (oder gar Beiträge unterschiedlicher Sprecher) verteilen.

[12] Eine Suche nach "(ein)?mal" über das Feld "Lemma" hätte denselben Effekt. Allerdings fügt die Lemmatisierung im Falle von Adverbien keine zusätzliche Information hinzu: normalisierte und lemmatisierte Formen sollten identisch sein. Die Wahl des Feldes "Normalisiert" ist daher insofern die bessere Option, als hierbei auch keine bei der automatischen Lemmatisierung eventuell eingeführten Fehler zum Tragen kommen können.

Abbildung 10: Kontextfilter mit regulärem Ausdruck "(ein)?mal"

In der KWIC werden nun alle Treffer von *sagen*, in denen *(ein)mal* nicht im Abstand von höchstens 15 Tokens im linken oder rechten Kontext vorkommt, als abgewählt markiert. Es verbleiben 937 ausgewählte Treffer, in denen die Form *(ein) mal* durch Fettdruck hervorgehoben ist. Wir löschen jetzt die abgewählten Suchergebnisse durch Klick auf das Papierkorb-Symbol und mischen die verbleibenden Ergebnisse zufällig durch . Letzteres dient dazu, die nun folgende partielle händische Analyse nicht dadurch zu verfälschen, dass nur wenige Interaktionen, die wegen der willkürlichen Reihenfolge der Interaktionen im Korpus zufällig am Anfang des Suchergebnisses stehen, berücksichtigt werden.

Abbildung 11: Kookkurrenzen von *sagen* und *mal*, zufällig durchmischt

4.2.2. Manuelles Bearbeiten der Suchergebnisse

Wir beginnen nun, die Suchergebnisse einzeln zu betrachten, um zu entscheiden, welche die uns interessierenden Konstruktionen belegen. Dabei gehen wir großzügig vor, d.h. sortieren zunächst nur solche Ergebnisse aus, in denen eindeutig *keine* diskursmarkerartige Verwendung vorliegt. Diese Treffer wählen wir ab, indem wir das Häkchen in der ersten Spalte der KWIC entfernen.

Abbildung 12: Manuelles Aussortieren von *false positives*

Dabei kann es vorkommen, dass die in der KWIC angezeigte Information für eine Entscheidung nicht ausreicht. Zum einen ist auf der Grundlage der transkribierten Formen alleine nicht immer klar, in welcher Beziehung *mal* und *sagen* zueinander stehen. In diesen Fällen kann über den Play-Button ▷ der zugehörige Ausschnitt der Aufnahme abgespielt werden und die somit verfügbare prosodische Form der Äußerung bei der Entscheidung mit berücksichtig werden. Zum anderen kann auch einfach der angezeigte Kontext für eine Interpretation nicht ausreichend sein, weil nicht klar ist, ob und wie die angezeigte mit vorherigen oder nachfolgenden Äußerungen in Zusammenhang steht. In diesen Fällen kann über den Button der zugehörige Transkriptausschnitt angezeigt werden.

Beispielsweise legt die Anzeige des folgenden Beispiels in der KWIC zunächst nahe, dass es sich um eine Imperativ-Form handelt, die folglich aus dem Ergebnis auszusortieren wäre (FOLK_E_00187_SE_01_T_02[13]: Sprachbiografisches Interview).

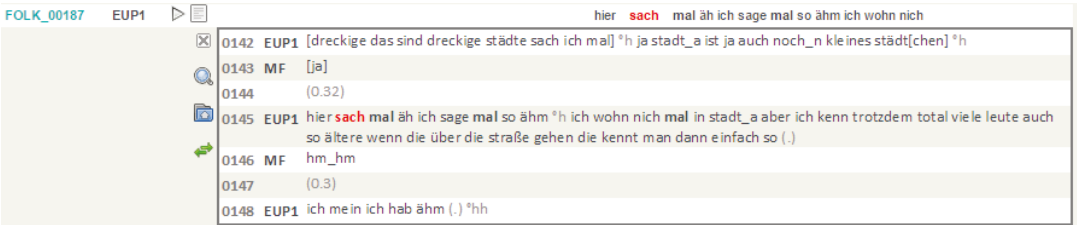

Abbildung 13: Transkriptkontext zu einem Treffer in der KWIC

Eine genauere Inspektion des Transkript-Ausschnitts zeigt aber, dass *sach mal* hier Teil einer Disfluenz ist, zu der auch der nachfolgende Häsitationsmarker sowie das Reparandum *ich sage mal* gehören. Darüber hinaus offenbart ein Anhören des zugehörigen Audios, dass der Sprecher an dieser Stelle sehr schnell spricht und die Entscheidung des Transkribenten, das Gehörte als "hier sach mal" zu verschriftlichen, zumindest nicht unstrittig ist, eine Transkription als "hier_sch sach mal" z.B. ebenso plausibel wäre. Entgegen des anhand der KWIC gewonnen Eindrucks muss dieser Treffer also nicht ausgefiltert werden.

Eine Analyse der ersten 100 (von 928) Treffern lässt vermuten, dass bei relevanten Konstruktionen die Tokens *mal* und *sagen* kaum (in meiner Stichprobe: nie) in einem Abstand von mehr als drei Tokens voneinander stehen. Dies erlaubt

[13] Diese Beispielreferenz sowie alle folgenden sind per Hyperlink mit dem entsprechenden Transkript-Ausschnitt in der DGD2 verknüpft. Registrierte Nutzer können so die Beispiele mit Ton durch Klick auf den Link aus der DGD2 abrufen.

uns, die Suche mit einem kleineren Kontextfenster von drei Tokens zu wiederholen und damit die Zahl der manuell zu inspizierenden Treffer deutlich (von 928 auf 589) zu reduzieren.

Diese Liste der verbleibenden 589 Kookkurrenzen von *mal* und *sagen* analysieren wir jetzt vollständig, indem wir zunächst alle eindeutig nicht relevanten Treffer aussortieren. Dazu gehören:

a) Fälle, in denen *sagen* in Verbindung mit *mal* in einem normalen Matrixsatz mit zugehörigem, über eine Subjunktion eingebettetem Nebensatz verwendet wird:

```
0199  BÄ    ja das sollte [man] aber das müsste man aber erst mal
            sich noch mal öhm (.) deutlich machen
            können sie jetzt noch mal sagen warum sich theater °h
0200  KS    [das]
0201        (0.47)
0202  BÄ    pädagogische ansätze oder auch handlungs und
            produktionsorientierte ansätze jetz besonders anbieten
```

[FOLK_E_00034_SE_01_T_01: Prüfungsgespräch an der Hochschule]

b) Weitere Verwendungen von *sagen* als *Verbum dicendi* ohne zugehörige Subjunktion:

```
1825  HM    also (.) von s es is net viel was er mache muss (.)
            hascht du mer mol so gsat
            abber ich denk abber so in seiner wahrnehmung is es
            wohl schon so ne
```

[FOLK_E_00026_SE_01_T_02: Meeting in einer sozialen Einrichtung]

c) Imperative:

```
0273  SM    also eine hast du genannt
0274        (1.14)
0275  SM    sag_s no_mal
0276        (0.51)
0277  CC    also (.) da wo er mit dem schachspieler redet
```

[FOLK_E_00121_SE_01_T_03: Unterrichtsstunde im Wirtschaftsgymnasium]

d) Fälle, in denen *sagen* und *mal* zu verschiedenen Konstruktionen oder Syntagmen gehören:

```
1476  AW    kann man der mutter net sagen sie soll mal [kurz zum]
            ohrenarzt gehen mit dem kin[d und ne]
1477  SZ    [ach nee]
1478  SZ    [nee stimmt]
```

[FOLK_E_00026_SE_01_T_03: Meeting in einer sozialen Einrichtung]

e) Fälle, in denen *ich sag mal* einen performativen Sprechakt einleitet. Dies ist ein Spezialfall, der nur in der Interaktion FOLK_E_00021, dort aber gehäuft (insgesamt 23 Belege), auftritt:

```
0663  DK    ja m m mach ja auch ich wart ja nur auf_n neuen °h (.)
            äh ich sag mal äh[m minta]l
0664  CH    [ich glaub nich dass der]
0665        (0.23)
0666  DK    im mittelfeld
```

[FOLK_E_00021_SE_01_T_12: Spielinteraktion zwischen Erwachsenen]

Ein Blick in die Metadaten des Gesprächsereignisses (Klick auf die Ereignis-Kennung in der KWIC) liefert u.a. folgende Informationen zu dieser Interaktion:

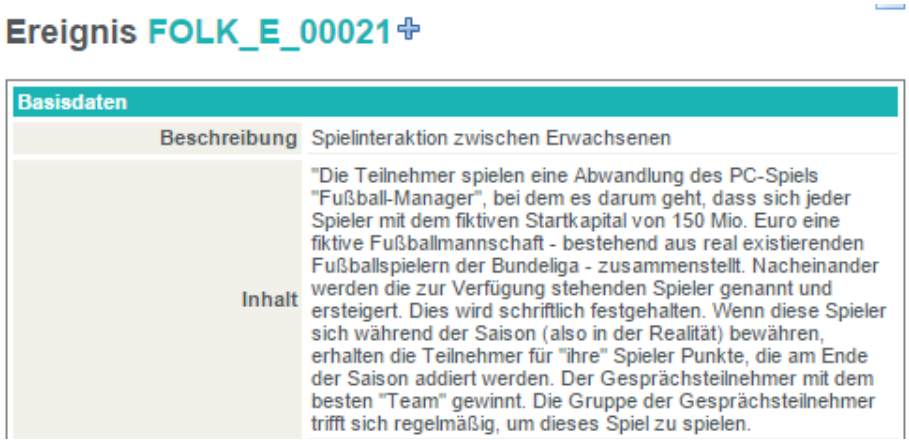

Abbildung 14: Gesprächsmetadaten

Die Formel *ich sag mal* leitet hier demnach den performativen Sprechakt ein, bei dem ein Spieler zur Versteigerung genannt wird.

Die vollständige Inspektion von 589 Suchergebnissen nimmt einige Zeit in Anspruch. Alternativ wäre auch denkbar, die Liste durch Ziehen einer Zufallsstichprobe (über den Button ⚄) zunächst geeignet (z.B. auf 100 Ergebnisse) zu reduzieren und in der Annahme, dass an einer ausreichend großen Zufallsstichprobe keine grundsätzlich anderen Phänomene festzustellen sein sollten als an der Gesamtheit der Ergebnisse, nur mit dieser Stichprobe weiterzuarbeiten. Für manche Untersuchungen stellt dies die einzig praktikable Möglichkeit dar, denn je nach untersuchtem Phänomen können Suchergebnisse schnell einen Umfang erreichen, der sich nicht mehr mit vertretbarem Aufwand manuell bearbeiten lässt.[14] Für die vorliegende Analyse inspizieren wir jedoch alle 589 Ergebnisse. Da wir dafür in der Regel mehr als eine Analysesitzung benötigen, speichern wir die bearbeitete Liste Ergebnisse am Ende jeder Sitzung über den Button 💾 unter einem geeigneten Namen und können sie dann jederzeit bei einer neuen Sitzung über den Button 📂 wieder aufrufen.

[14] Man denke etwa an eine Anfrage nach dem Häsitationsmarker *äh*, der sich in der aktuellen FOLK-Version 21.205 mal belegen lässt. Die DGD liefert bei allen Anfragen mit mehr als 10.000 Ergebnissen grundsätzlich nur eine Zufallsstichprobe als Ergebnis zurück, weil größere Ergebnisse auch die Performanz des Systems beeinträchtigen würden.

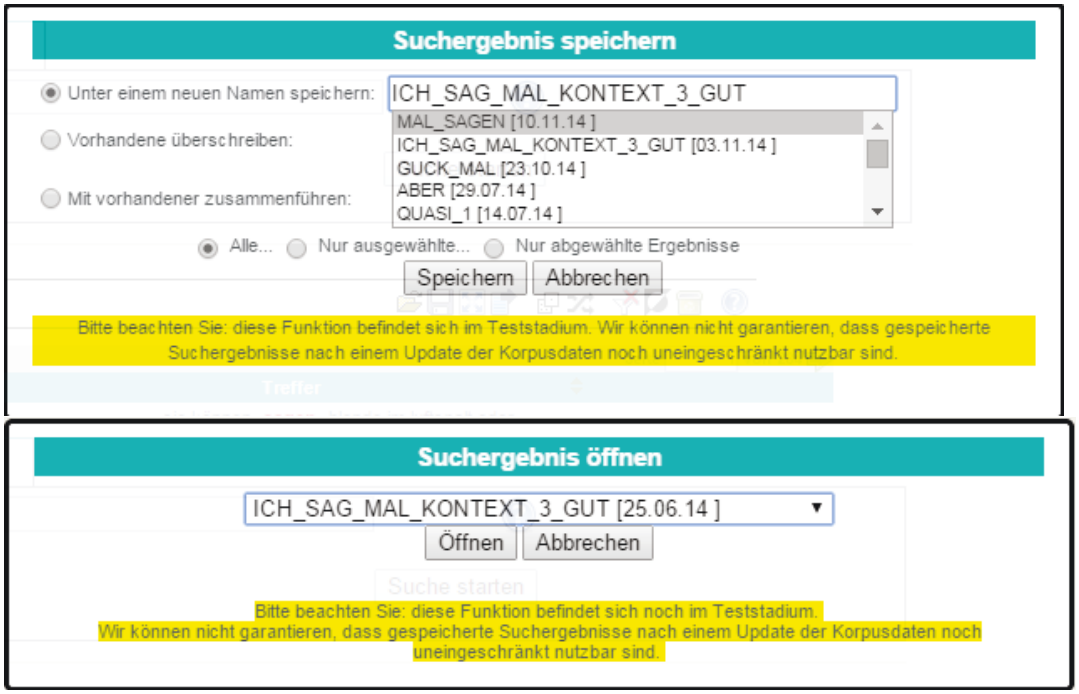

Abbildung 15: Speichern und Öffnen von Suchergebnissen

Auf diese Weise lassen sich 214 Suchergebnisse eindeutig aussortieren, so dass insgesamt 275 Belege verbleiben,[15] die mit der Konstruktion *ich sag mal* in Verbindung stehen und auf die wir die nun folgenden Analysen basieren.

4.2.3. Analyse der Belege

Die verbleibenden 275 Belege sollen nun im Hinblick auf die eingangs gestellten Fragen nach der Variation von Form und Funktion analysiert werden. Eine mögliche Herangehensweise ist, die Belege zunächst nach ihrem rechten Kontext zu sortieren (Klick auf das Symbol ⇳ im betreffenden Spaltenkopf), so dass Regelmäßigkeiten in den auf *sagen* folgenden Tokens sichtbar werden. Da es hier zunächst um einen Überblick über die Gesamtheit der Belege geht, mag es hilfreich sein, diese vollständig anzuzeigen. Wir lassen uns also über den Button  eine Vollansicht der KWIC anzeigen. Alternativ kann das gesamte Ergebnis über den Button  auch in eine Textdatei exportiert werden, die dann wiederum zur

[15] Die exakten Zahlen sollen hier aber keine Rolle spielen. Ginge es um eine belastbare Quantifizierung, müsste erstens der Prozess des Aussortierens expliziter beschrieben und methodisch abgesichert werden, also z.B. überprüft werden, inwieweit verschiedene Personen auf der Grundlage gleicher Regeln (wie hier in a) bis e) skizziert) zu den gleichen Entscheidungen kommen (*inter-annotator agreement*). Zweitens müsste geklärt werden, mit welcher Begründung das Korpus überhaupt als Grundlage verwendet werden kann, um aussagekräftige Schlüsse über Einzelfrequenzen aus einem Suchergebnis abzuleiten (Repräsentativität). Dies ist hier nicht das Ziel, denn es ist zweifelhaft, ob FOLK in seinem jetzigen Umfang überhaupt für solche quantitativen Analysen geeignet ist. Es bleibt aber in jedem Fall plausibel, auf der Grundlage von FOLK in seiner aktuellen Form die Existenz von Formen und Mustern zu konstatieren und Tendenzen zu deren Häufigkeit zu beschreiben. In diesem Sinne sind die im folgenden Abschnitt genannten Zahlen zu verstehen.

weiteren Bearbeitung in Tabellenkalkulationen wie Excel eingelesen werden kann.

Ergebnisse 1 bis 275 von 275 (275 / 0 aus-/abgewählt)

	Ereignis	Sprecher	Treffer		
1	FOLK_00001	RG	isch	sach	**ma** die geberspannung dann messe also fünf volt oder so
2	FOLK_00004	JM	ich würde jetz **ma**	soche	fochlich un menschlich
3	FOLK_00004	GS	jetz wollte ma ja ich	sach	**ma** unabhängig von fachkompetenze des beurteile
4	FOLK_00004	GS	der	sagen	wa **mal** wir sind ein äh wir haben mehrere zugpferde
5	FOLK_00004	ML	sein dass ne ganze abteilung dann äh	sagen	wa **ma** in die roten zahlen reinrutscht
6	FOLK_00004	TF	nur für mich	sach	ich jetz **ma**
7	FOLK_00004	GS	wie zum beispiel wenn ich ich	sag	jetz **mal** n schweiß
8	FOLK_00004	ML	wie gesagt gruppenarbeit und danach	sag	ma **ma** wemma
9	FOLK_00005	ML	wenn	sagen	ma **mal**
10	FOLK_00006	AK	also ich würd **mal**	sage	also s so lang der
11	FOLK_00006	LB	ja so n puchsignal	sach	isch **mol** sieht immer annerschd aus wie die realität
12	FOLK_00006	RG	net die unterbreschung vom schirm wenn des	sa	**mol** des kabel irgendwie gequetscht is dann und s kummt
13	FOLK_00006	RG		sag	mer **mol** ++++++++++++
14	FOLK_00007	GS	ja ich ich	sag	s e mol so ketzerisch unter anderem aber anscheinend is
15	FOLK_00007	JK		sag	ich **mal** wenn wir sehen dass dass irgendjemand was uns
16	FOLK_00007	GS	fähigkeiten haben ich	sach	**ma** so
17	FOLK_00007	JK	ich	sag	**mal**
18	FOLK_00007	RZ	grundgleichung ich	sag	**mal** so
19	FOLK_00007	RG	ich	sag	**mal** chemisch mir ist es ab und zu mal ufgfalle
20	FOLK_00007	ML	über die kreuzung fährt ähm und dort	sagen	wir **mal** schon zwei autos sind dass es eventuell net
21	FOLK_00007	GS	ja s könnten se im einführungsgschpräch	sa	ma **ma** mach mer kurz ja
22	FOLK_00009	TF	könnt ich ja jetz kein fünfzylinder bauen	sag	ich **mal**
23	FOLK_00015	CH	worden deshalb ist das wichtig i ich	sag	s **mal** so äh
24	FOLK_00015	CH		sag	ich jetz **mal** ein soundfile haben und dann eben eine
25	FOLK_00015	CH	einem etwas etwas größeren bereich also sie	sagen	jetz **mal**
26	FOLK_00015	CH	i ich	sag	jetz **mal**
27	FOLK_00015	CH	ich sa	sag	jetz **mal** äh westmitteldeutschen raum viel mehr
28	FOLK_00015	CH	ich	sag	jetz **mal** so ba ganz banal ich s s nehm
29	FOLK_00015	CH	ich	sa	sag jetz **mal** äh westmitteldeutschen raum viel mehr
30	FOLK_00018	EM	also	sa	ma **mal** so nee sa ma mal so ich hab
31	FOLK_00018	EM	also sa ma **mal** so nee	sa	ma **mal** so ich hab jetz also an nem montag
32	FOLK_00018	EM	hm ratatouille wurde aber auch goldig gemacht	sa	ma **mal** so ne
33	FOLK_00020	HM	macht dass er net irgendwie ha ich	sag	**mal**
34	FOLK_00022	HM	entscheidung zu treffe oder selwer **mal** zu	sage	
35	FOLK_00022	NG	nee umsonst bin ich nich da	sagen	wir s **mal** so weil
36	FOLK_00024	MS	also ich	sach	jetz **mal** was ganz ehrlich es gibt für uns
37	FOLK_00024	MS	mitmacht	sag	ich **mal** ne
38	FOLK_00024	MS	also ich	sag	dir jetzt **mal** ganz ehrlich was ich hab m johannes
39	FOLK_00026	MS	des auch zu viel für n hobby	sag	ich **mal**
40	FOLK_00026	MS	ja is ja dann ach entpannend	sach	ich **mal**
41	FOLK_00026	MS	gesagt hast des auch net äh erzwingen	sag	isch **mal** aber
42	FOLK_00027	AM	is da besser trotzdem struk äh strukturiert	sag	ich **mal** in der hinsicht und äh und diese strukturierung
43	FOLK_00028	FF	wissenschaftler die sich in dem äh hm	sag	ich **mal** in den siebziger jahren des neunzehnten jahrhunde…
44	FOLK_00028	FF	öhm kulturentwicklung beeinflusst is und öh sich	sag	ich **mal** gesetzeswissenschaftsaspekte zum beispiel wie di…
45	FOLK_00028	FF		sag	ich **mal** in den fokus gerückt weil sie halt auch
46	FOLK_00028	FF	dass diese äh dä dass dass ähm	sag	ich **mal** das phonem noch nich bedeutungs
47	FOLK_00030	PB		sagen	mer **ma** dreihundertfünfzig
48	FOLK_00031	CR	ja	sach	**mal** passive wissensaufnahme

Abbildung 16: Vollansicht der KWIC-Konkordanz

Wir halten zunächst zwei offensichtliche Befunde fest:

f) Die Form *ich sag mal* (also die "Grundform" ohne Einschübe oder morpho-syntaktische Variation) ist insgesamt 42-mal belegt. 5 dieser 42 Belege haben die Form *ich sag mal so*.[16]
g) Die invertierte Form *sag ich mal* ist sogar noch häufiger (75-mal) belegt.

[16] Auch für solche Auszählungen können Kontext-Filter und/oder die Sortierfunktion für die KWIC benutzt werden, um eine Vorauswahl an relevanten Belegen zu treffen.

4.2.4. *ich sag mal* als abschwächender Diskursmarker

Wir konzentrieren uns nun zunächst auf diejenigen Belege, die sich von Form und Funktion her mit der Analyse von Auer/Günthner (2003) in Einklang bringen lassen:

h) Es ist ohne weiteres möglich, in den 42 Belegen der Grundform solche zu finden, in denen *ich sag mal* eindeutig als Diskursmarker im Sinne von Auer/Günthner (2003) verwendet wird. So verwendet der Sprecher ZIT4 im folgenden Beispiel *ich sag mal*, um die folgende Aussage ("ich schreibe besser als andere") zu markieren und seinem Gesprächspartner zu signalisieren, dass er sich deren potentiell problematischen Status (Selbstlob) bewusst ist. Die Funktion von *ich sag mal* ist hier also mit "abschwächend" treffend beschrieben.

```
0653  ZIT4  [na wenn man so n ich] will ni sagen hochtrabend
            weil ich schreib eigentlich ne hochtrabend
            aber °h ich sag ma ich schreib vielleicht
            in eenen bessren stil als irgendwie
0654        (0.49)
0655  ZIT4  manche anderen leute ich schreib och in en
            schlechteren an wie als wiederum andere […]
```

[FOLK_E_00179_SE_01_T_01: Sprachbiografisches Interview]

i) Auch für die invertierte Form lässt sich die von Auer/Günthner (2003) beschriebene Diskursmarkerfunktion eindeutig nachweisen. So bezieht sich *sag ich mal* im folgenden Beispiel auf ein vorangehendes Zugeständnis der Sprecherin MS (sie möchte die Teilnahme einer Schülerin an einer Klassenfahrt nicht erzwingen), wobei in den nachfolgenden Äußerungen aber wiederum Vorbehalte gegen dieses Zugeständnis zum Ausdruck gebracht werden (die für die Nicht-Teilnahme angeführten Gründe überzeugen die Sprecherin nicht). *sag ich mal* schwächt hier also das Zugeständnis mit Blick auf die verbleibenden Vorbehalte ab.

```
1868  MS    die argumente versteh isch auch
            und isch denk wenn sie sich jetzt wirklich
            mit händ und füß äh (.) wehrt dagegen
1869        (0.73)
1870  MS    auf die klassenfahrt (.) zu gehen
1871        (0.48)
1872  MS    ja (.) äh (.) würd isch aus den gründen die du jetzt
            gesagt hast des auch net (.) äh erzwingen sag isch mal
            °hh aber
1873        (0.33)
1874  MS    isch muss ehrlich sagen
            des was du jetzt sagst was die frau bach gesagt hat
1875        (0.76)
1876  MS    sind für mich (.) äh keine gründe
            sie is noch wie ein baby zu hause
            °h da nimmt sie ihr schnuffel[tuch mit
            meinste ernst]haft ich bin ohne mein stofftier auf
            klassen[fahrt gef]ahren
```

[FOLK_E_00026_SE_01_T_03: Meeting in einer sozialen Einrichtung]

j) Die naheliegende Vermutung, dass – wie in den Beispielen in h) und i) – *ich sag mal* sich generell auf Nachfolgendes, *sag ich mal* hingegen generell auf Vorhergehendes bezieht, lässt sich nicht bestätigen. Insbesondere finden sich beide Formen auch in Einschüben an Positionen, die nicht "syntaktisch peripher" (Auer/Günthner 2003, s.o.) sind :

```
0871  GS    jetz wollte_ma ja ich sach ma unabhängig von
            fachkompetenze des (.) beurteile
```

[FOLK_E_00004_SE_01_T_01: Unterrichtsstunde in der Berufsschule]

```
0534  GH    wir stellen sie infrage
0535        (0.42)
0536  GH    ob sie die richtigen prioritäten setzt und die
            effizienteste lösung (.) für unser problem is
            wir brauchen eine (.) beschleunigung °hh in (.) der
            achse (.) ich sag mal mannheim münchen
```

[FOLK_E_00064_SE_01_T_05: Schlichtungsgespräch]

```
0137  WL    °h des hat aber hier
            mit stuttgart einundzwanzig sag i mal nur insofern zu
            tun (.) als eben (.) unter äh äh (.) ähm
            ((räuspert sich)) äh s untersuchungen da (.) oder (.)
            baumaßnahmen da waren (.) °h die bereits tiefer
            eingeschnitten haben
```

[FOLK_E_00069_SE_01_T_05: Schlichtungsgespräch]

```
0009  FF    alles klar °hh also die junggrammatiker is öhm (.) ähm
            eine (.) gruppe junger wissenschaftler
            die sich in dem äh hm sag ich mal °h in den siebziger
            jahren des neunzehnten jahrhunderts äh °h
            sozusagen ihren öh pf einführung hatte [...]
```

[FOLK_E_00028_SE_01_T_01: Prüfungsgespräch in der Hochschule]

k) Neben nachfolgendem *so* (s.o.) lassen sich wie erwartet auch *jetzt* und *einfach* als eingeschobene Begleiter der Konstruktion belegen, wobei *jetzt* deutlich häufiger (41-mal) auftritt als *einfach* (5-mal)

```
0129  TF    weil ich mach_s ja im grund net
0130        (0.39)
0131  TF    nur für mich sach ich jetz ma (.)
```

[FOLK_E_00004_SE_01_T_02: Unterrichtsstunde in der Berufsschule]

```
0082  PB_aa (.) ja vom papier her
0083        (0.31)
0084  PB_aa [ja] aber in ihrem kopf sag ich einfach mal war_s
            noch nich vorbei das ham_se ja vorhin n paarmal gesagt
```

[FOLK_E_00160_SE_01_T_02: Gespräch im Polizeirevier]

l) Darüber hinaus finden sich bei mehreren Belegen adverbiale Begleiter, die die durch *mal sagen* geleistete Abschwächung weiter qualifizieren bzw. motivieren.

```
0802  MF    besonders gemacht hat °h (.) und ((knarrt)) ja der
            orient da (.) is eben das wo dann wo die juden auch
0803        (0.4)
0804  MF    herkommen sag ich jetz mal so (.) pauschalisiert ne °h
            un deswegen äh hat sie da ne verbundenheit gefühlt
            [...]
```

[FOLK_E_00059_SE_01_T_01: Prüfungsgespräch in der Hochschule]

Neben *pauschalisiert* finden wir hier zum einen Adverbien wie *ganz allgemein*, *grob*, *verkürzend*, *banal*, *plump* und *ketzerisch*, die jeweils eine potentiell problematische Eigenschaft der betreffenden Aussage spezifizieren, die eine Abschwächung angemessen erscheinen lässt, zum anderen aber auch *ungeschützt* und *vorsichtig*, die direkt die Haltung des Sprechers zur betreffenden Aussage bzw. zu deren Abschwächung thematisieren.

m) Insbesondere in den Transkriptionen der Maptask-Dialoge markiert *ich sag mal* oft, dass der Sprecher eine Aussage als eventuell nicht hinreichend präzise empfindet (insgesamt 18 Belege). Im folgenden Beispiel folgt dann auch eine nachträgliche Paraphrasierung, die die Ungenauigkeit des ersten Formulierungsversuchs zu beheben versucht. Die Funktion der Abschwächung ist hier sicherlich weiterhin gegeben, sie motiviert sich aber anders als in den meisten bisher angeführten Beispielen nicht daraus, dass der markierte Ausdruck eine (z.B. "politisch inkorrekte") Evaluation implizieren kann, von der sich der Sprecher mit seiner Abschwächung teilweise wieder distanziert.

```
0728  TMP4  °h und dann machste so_n (.) ich sach mal
            so_n viertelkreis is des
0729        (0.24)
0730  TMP4  n viertelkreis (.) dass de halt nicht so_ne
            spitze ecke jetzt hast
```

[FOLK_E_00091_SE_01_T_01: Maptask]

4.2.5. Andere Formen von *mal sagen*

Wir haben uns bis hier auf die Form *ich sag mal* und deren invertierte Variante *sag ich mal*, jeweils mit möglichen adverbialen Einschüben, konzentriert. Diese machen insgesamt aber nur etwas über die Hälfte der Gesamtbelege aus. Hinzu kommen noch folgende andere Kookkurrenzen von *mal* und *sagen*:

n) Häufig (24-mal) findet sich in den Belegen die Konjunktiv-Variante *würde ich mal sagen*. Unter den Verwendungsweisen sind solche wie die folgenden, die sich hinsichtlich ihrer Funktion (Abschwächung des negativ konnotierten Ausdrucks *dickerer Mensch* bzw. Markierung der unpräzisen Angabe *hundertzwanzig*) kaum von der Indikativ-Variante unterscheiden lassen.

0556 JN wenn er sich dann da durchzwängt un un_e platz sucht
[wie wenn er] die ganze fahrt lang in dem gedränge
steht jetzt als dickerer mensch
würd ich jetzt mal sage

[FOLK_E_00121_SE_01_T_03: Unterrichtsstunde im Wirtschaftsgymnasium]

0595 BFD3 ähm (.) wie viel is das ((lacht))
so_n bisschen (.) äh mehr als neunzig grad (.)
so hundertzwanzig **würd ich mal sagen**

[FOLK_E_00109_SE_01_T_01: Maptask]

o) Noch häufiger (80 Belege) findet sich die Variante mit der ersten Person Plural *sagen wir mal*. Auch hier sind ohne weiteres Verwendungsweisen zu belegen, deren Funktion sehr ähnlich zu den Varianten in der ersten Person Plural (distanzierende Abschwächung zum Bezug auf Nacktheit im ersten, Markieren einer Angabe als unpräzise im zweiten Beispiel) ist.

1200 DG die frau rechts hinter ihm
1201 (0.8)
1202 DG isch die sattelnase
1203 ((allg. Gekicher, 2.39s))
1204 DG und
1205 DG sie hat halt ja (.) **sa_ma_mal** en hauch von nichts an
1206 (0.42)
1207 DG un
1208 (0.58)
1209 PM die hat schon was an
1210 DG ja [aber halt n bisschen]
1211 PM [is doch geil o]der (.) oder

[FOLK_E_00124_SE_01_T_01: Unterrichtsstunde im Wirtschaftsgymnasium]

0102 FFM4 °h dann gehst jetzt einfach runter also dann ers
kommst du erst an den spiegeln vorbei und dann am rad
0103 (0.24)
0104 FFM3 j[a]
0105 FFM4 [°hh] aber nicht zu weit runter und dann **sag mer mal**
bisschen bisschen weiter runter bis als die räder sind
°h gehst du weiter (.) nach links

[FOLK_E_00089_SE_01_T_01: Maptask]

Teilweise sind diese Belege treffender als Ankündigung eines Beispiels charakterisiert:

0884 CS oder ja (.) wenn wenn
0885 (0.34)
0886 CS des unternehmen an die börse geht so °h und dann (.)
kommt_s zu_nem nennbetrag **sage_ma jetz ma** zwanzig euro
0887 (0.91)
0888 CS sprich ich zahl zwanzig euro (.) für die aktie

[FOLK_E_00166_SE_01_T_01: Unterrichtsstunde im Wirtschaftsgymnasium]

Die nicht-invertierte Form in der ersten Person Plural (*wir sagen mal*) ist hingegen nicht belegt.

p) Schließlich enthält die Belegsammlung auch einige wenige Beispiele (5 Belege) mit der Partizipial-Form *mal gesagt*, die wiederum funktional kaum von den anderen Formen unterschieden werden können.

```
0110  AP    ja (.) °h ähm also (.) die abgrenzung ist
            erst ma grob gesagt (.) phraseologismen sin feste
            wortverbindungen die aus mindestens zwei lexemen
            bestehen [also] polylexikal sind °h und (.) ähm
```

[FOLK_E_00056_SE_01_T_01: Prüfungsgespräch in der Hochschule]

```
0302  KA    herr professor lächler °h [°h ] äh des sieht so aus
            (.) jetzt laienhaft mal gesagt (.) der bahnhof liegt
            im (.) grundwasser so drin (.) wie ein tanker (.) äh
            auf see
```

[FOLK_E_00069_SE_01_T_05: Schlichtungsgespräch]

4.2.6. Abhängigkeiten von Gesprächs- und Sprechertypen

Zu den von den Studierenden geäußerten Intuitionen (s.o.) gehörten auch verschiedene Vermutungen, dass die Verwendung von *ich sag mal* auf bestimmte Gesprächs- oder Sprechertypen beschränkt ist oder bei diesen zumindest gehäuft auftritt. Nicht alle diese Intuitionen können anhand von FOLK überprüft werden, da das Korpus in seiner aktuellen Version nicht annähernd alle potentiell interessierenden Gesprächs- oder Sprechertypen (z.B. keine Interviews mit Profi-Fußballern, keine Talkshows) enthält. Anhand der in der DGD verzeichneten Metadaten können wir aber immerhin einige solcher Abhängigkeiten prüfen.

Wir tun dies zunächst exemplarisch für die regionale Verteilung unserer 275 Belege. Dazu wechseln wir in den Reiter METADATEN. Dort wählen wir als ersten Deskriptor "E: Ort (Region)" aus. Das "E" signalisiert dabei, dass es sich um eine Eigenschaft des Ereignisses handelt, unter diesem Deskriptor also der Ort vermerkt ist, an dem ein Gespräch stattgefunden hat. Mit einem Klick auf das Symbol ✚ fügen wir einen zweiten Deskriptor "S: Aufenthaltsorte (Region)" hinzu. Hier signalisiert das "S", dass es sich um eine Eigenschaft des Sprechers handelt.

Abbildung 17: Einblenden ausgewählter Metadaten in der KWIC-Konkordanz

Über die Schaltfläche METADATEN ANZEIGEN / FILTER ANWENDEN[17] blenden wir dann in zwei zusätzlichen Spalten der KWIC für jedes Suchergebnis die zugehörigen Metadaten ein:

Abbildung 18: KWIC-Konkordanz mit Metadaten

Bereits auf den ersten Blick lässt sich sehen, dass die Einträge in den hinzugefügten Spalten stark variieren. Durch Klick auf den betreffenden Spaltenkopf können wir die Spalten sortieren und manuell[18] auszählen:

Sprachregion	Ort (Gespräch)	Aufenthaltsorte (Sprecher)[19]
Rheinfränkisch	47	72
Obersächsisch	47	13
Alemannisch	43	19
Schwäbisch	31	42
Ripuarisch	27	28
Nordniederdeutsch	20	20
Hessisch	19	28
Brandenburgisch	13	13
Thüringisch	6	5
Bairisch	3	5
Moselfränkisch	2	5
Westfälisch	1	1

Tabelle 2: Belege nach Sprachregionen (absolute Zahlen)

Die Auszählung bestätigt, dass unsere Belege sich über den gesamten deutschen Sprachraum verteilen, unabhängig davon, ob man den Ort des Gesprächs oder die Aufenthaltsorte des Sprechers zugrunde legt.

Mit der Beobachtung, dass die Belege sich *nicht gleichmäßig* über den deutschen Sprachraum verteilen – beispielsweise haben wir gar keine Belege aus der ostfränkischen Sprachregion, und die Belegzahlen etwa für die Bairische und die Rheinfränkische Sprachregion unterscheiden sich erheblich – ist hingegen äußerst zurückhaltend umzugehen. Insbesondere müssen wir bedenken, dass FOLK insgesamt nicht regional ausgewogen ist, dass daher nicht für jede Sprachregion gleiche Mengen an Daten durchsucht wurden, und dass die absoluten Häufigkeiten sich somit nicht einfach in entsprechende relative Häufigkeiten übersetzen lassen. Prinzipiell plausiblere Vergleichswerte erhalten wir, wenn wir die Anzahl der

[17] Wir nutzen den Reiter hier nur zum *Anzeigen* von Metadaten. Ein *Filter* würde zusätzlich für jeden Deskriptor einen oder mehrere Werte im rechts daneben stehenden Textfeld spezifizieren. So könnten z.B. nur diejenigen Suchergebnisse ausgewählt werden, deren "Ort (Region)" als "Obersächsische Sprachregion" dokumentiert ist. Dies kann z.B. für die folgenden Auszählungen genutzt werden.

[18] Funktionalität, die ein solches Auszählen automatisch unterstützt, befindet sich in Planung.

[19] In den Metadaten zu FOLK-Sprechern sind nach Möglichkeit alle Orte festgehalten, in denen sich ein Sprecher im Laufe seines Lebens längere Zeit aufgehalten hat. Dies erklärt, warum sich die Aufenthaltsorte zu einer größeren Zahl aufaddieren als die Gesprächsorte.

Belegstellen pro Region ins Verhältnis setzen mit der Anzahl der pro Region durchsuchten transkribierten Tokens:

Sprachregion	Ort (Gespräch)	Rang absolut	Durchsuchte Tokens	Verhältnis
Ripuarisch	27	5	24822	0.10877%
Brandenburgisch	13	8	16386	0.07934%
Westfälisch	1	12	1545	0.06472%
Alemannisch	43	3	81297	0.05289%
Thüringisch	6	9	13934	0.04306%
Nordniederdeutsch	20	6	49611	0.04031%
Obersächsisch	47	1	127190	0.03695%
Schwäbisch	31	4	118119	0.02624%
Moselfränkisch	2	11	8986	0.02226%
Hessisch	19	7	110359	0.01722%
Rheinfränkisch	47	1	310023	0.01516%
Bairisch	3	10	54977	0.00546%

Tabelle 3: Belege nach Sprachregionen im Verhältnis zu durchsuchten Tokens

Auch aus dieser Auszählung – gemäß derer beispielsweise die Belege aus der rheinfränkischen Sprachregion relativ fast am seltensten sind, während sie bei den absoluten Zahlen den ersten Rang belegen (umgekehrt etwa beim einzigen Beleg aus der Westfälischen Sprachregion) – lassen sich aber keine weiterführenden Schlüsse ableiten. Zu beachten wäre beispielsweise außerdem, dass in FOLK die Gesprächstypen teilweise (aber nicht durchgängig) systematisch mit der Sprachregion zusammenhängen. So stammen alle Prüfungsgespräche aus der Obersächsischen Sprachregion, und es wurden in dieser Region bislang nur wenige andere Gesprächstypen erhoben.

Wir belassen es also bei der Beobachtung, dass unsere Belegsammlung keine offensichtlichen Tendenzen zur Verwendung von "mal sagen" in bestimmten Sprachregionen aufzeigt, sondern dass die Konstruktion vielmehr flächendeckend nachweisbar ist.

Zu ähnlichen Schlüssen gelangen wir, wenn wir uns andere Gesprächs- oder Sprechereigenschaften anschauen.

Wie schon die Liste der bis hier angeführten Beispiele andeutet, gibt es z.B. keine offensichtliche Beschränkung der Verwendung von *mal sagen* auf bestimmte Gesprächstypen. Wir können die Konstruktion in verschiedenen Typen privater Alltagsgespräche (Gespräch auf der Urlaubsreise, Gespräch in der Familie, Gespräch unter Freunden, Paargespräch, Studentisches Alltagsgespräch, Tischgespräch), verschiedenen institutionellen Gesprächsformen (Gespräch im Polizeirevier, Meeting in einer sozialen Einrichtung, Prüfungsgespräch in der Hochschule, Schichtübergabe in einem Krankenhaus, Unterrichtsstunde im Wirtschaftsgymnasium bzw. in der Berufsschule) sowie in einer öffentlichen Interaktion (Schlichtungsgespräch) nachweisen, außerdem auch in elizitierten, nur quasispontanen Gesprächen (Maptasks, biografische Interviews).

Desgleichen verfügen wir über Belege von männlichen wie von weiblichen Sprechern, sowie von Sprechern aller Altersstufen zwischen 16 und 80 Jahren.

Insgesamt können wir also festhalten, dass die Verwendung von *mal sagen* nicht an bestimmte Gesprächs- oder Sprechertypen gebunden zu sein scheint.[20] Wir können im Gegenteil davon sprechen, dass die Konstruktion in deutschen Gesprächen allgegenwärtig ist.

Dafür, dass es mithin nicht gerechtfertigt ist, die Verwendung der Konstruktion als "gelegentlich" (Auer/Günthner 2003, s.o.[21]) zu charakterisieren, spricht im Übrigen auch, dass sich die Anzahl der gefundenen Belege (275, s.o.) in der gleichen Größenordnung bewegt wie die Lemmafrequenz der Einzellexeme *Mama*, *Straße*, *manchmal* oder *normal*, die mit Sicherheit in den meisten gängigen Definitionen zum Grundwortschatz des Deutschen gezählt werden. Dies zeigt ein Blick in die Lemmafrequenzliste zu FOLK, die sich über **KORPORA > ZUSATZMATERIALIEN** abrufen lässt.

```
total 279
Mama 277
spät 277
Straße 277
Wort 276
ohne 275
lieb 274
irgendwo 273
manchmal 273
trotzdem 272
normal 269
```

Abbildung 19: Ausschnitt aus der Lemmafrequenzliste zu FOLK

4.3. *mal sagen* in anderen Korpora

Als Korpusplattform des Archivs für Gesprochenes Deutsch enthält die DGD neben FOLK noch viele weitere mündliche Korpora. Diese sind bezüglich ihres Designs, der erhobenen Gesprächstypen und der Art ihrer Erschließung sehr heterogen. Der größere Teil besteht etwa aus narrativen Interviews, die im Rahmen variationslinguistischer Erhebungen durchgeführt und "nur" standardorthographisch transkribiert wurden, und die sich mithin von authentischen gesprächsanalytischen Daten stark unterscheiden. Wenn man über FOLK hinaus weitere DGD-Korpora in eine Untersuchung einbezieht, ist daher noch einmal erhöhte methodische Vorsicht geboten. Dennoch kann eine Erweiterung der Datenbasis auf andere Teile des AGD-Bestandes nützlich und interessant sein, insbesondere weil sich mit den bis in die 1950er Jahre zurückreichenden Beständen auch eine diachrone

[20] Dies schließt nicht aus, dass die Konstruktion in bestimmten Gesprächstypen, in bestimmten Sequenztypen oder bei Sprechern eines bestimmten Typs signifikant häufiger auftritt als bei anderen. Wir stellen nur fest, dass es keine absolute Beschränkung auf einige wenige solcher Typen zu geben scheint und FOLK für eine weitergehende quantitative Analyse (noch) keine geeignete Datenbasis darstellt.

[21] Allerdings müssen dabei auch der Entstehungszeitpunkt der Arbeit von Auer/Günthner (2003) bzw. die Herkunft der dort verwendeten Daten beachtet werden. Wenn – wie die Ausführungen im folgenden Abschnitt nahelegen – *ich sag mal* sich erst nach der Wende in Westdeutschland zu verbreiten begonnen hat, mag die Frequenz des Phänomens in westdeutschen Gesprächen vor 2003 mit "gelegentlich" sehr treffend beschrieben sein.

Perspektive auf mündliche Sprache eröffnet. Wir demonstrieren dies exemplarisch, indem wir eine einfache Suche nach Konstruktionen mit *mal sagen* auf weiteren DGD-Korpora ausführen. Um den Aufwand in Grenzen zu halten, beschränken wir uns dabei auf die "kanonische" Form *ich sag mal*.

4.3.1. Korpora vor 1980

Wir wählen zunächst nur diejenigen Korpora aus, die vor 1980 erhoben wurden und keine auslandsdeutschen Varietäten enthalten. Es sind dies die Variations-Korpora "Deutsche Mundarten" (Zwirner-Korpus, ZW), "Deutsche Umgangssprache" (Pfeffer-Korpus, PF) und "Deutsche Mundarten: ehemalige deutsche Ostgebiete" (OS)[22] sowie die Gesprächskorpora "Grundstrukturen" (Freiburger Korpus, FR) und "Dialogstrukturen" (DS). Insgesamt umfassen diese fünf Korpora knapp 6 Millionen transkribierte Tokens (siehe ÜBER DIE DGD > BESTAND).

Eine Anfrage auf allen Korpora gleichzeitig ist prinzipiell möglich, liefert uns aber für das Ausgangslemma *sagen* 31.680 Treffer, die, wie oben beschrieben, automatisch auf eine Zufallsstichprobe von 10.000 Treffern reduziert werden. Um wirklich alle Belege für *ich sag mal* zu erhalten, müssen wir also entweder Korpus für Korpus vorgehen, oder die initiale Anfrage spezifischer machen. Wir entscheiden uns für letzteren Weg, wählen in der Korpusliste links die genannten fünf Korpora aus und suchen zunächst nach dem regulären Ausdruck "sag(e)?" (= *sag* oder *sage*) in den normalisierten Formen. Wir erhalten 1749 Treffer, die wir weiter reduzieren, indem wir im Reiter KONTEXT zunächst die Form *ich* im linken Kontext und dann die Form *mal* im rechten Kontext (Abstand jeweils ein Token) suchen.

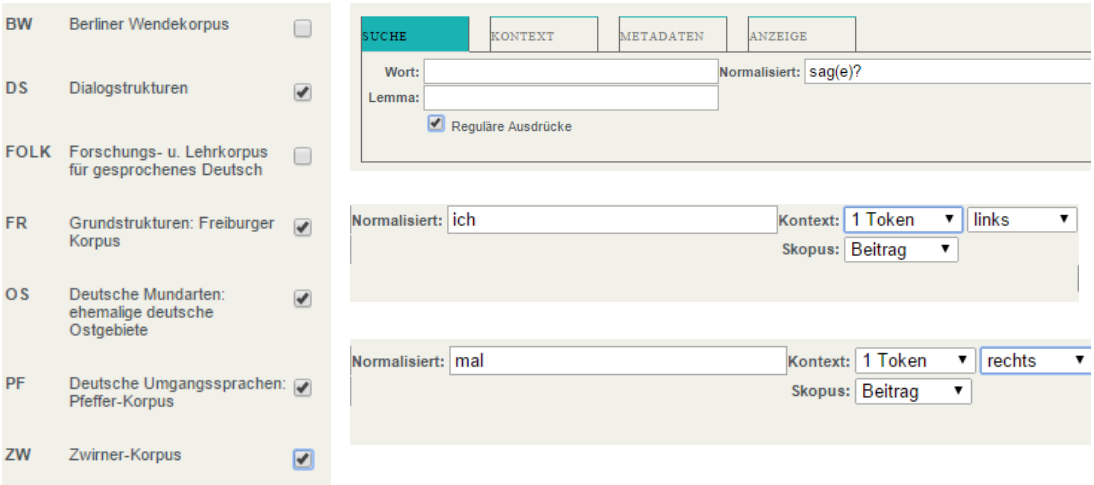

Abbildung 20: Token-Suche nach *ich sag mal* in DS, FR, OS, PF und ZW

Das Ergebnis ist durchaus überraschend: es verbleiben keine Suchergebnisse; die Form *ich sag(e) mal* lässt sich in den 6 Millionen transkribierten Wörtern mit anderen Worten kein einziges Mal nachweisen. Auch eine Suche nach der invertier-

[22] Die Definition trifft auch noch auf weitere DGD-Korpora wie z.B. SV und SW (Mundarten in Südewestdeutschland/Vorarlberg bzw. im Schwarzwald) zu. Allerdings liegen für diese Korpora keine Transkripte vor, so dass wir sie nicht in die Recherche einbeziehen können.

ten Form *sag ich mal* (ein einziger Treffer, bei dem es sich aber um eine normale *verbum dicendi*-Verwendung handelt) oder die Erweiterung des Kontexts auf 3 Tokens (11 Treffer, darunter gleichfalls keine diskursmarkerartigen Verwendungen) bleiben ergebnislos. Anders verhält es sich mit der Variante *sagen wir mal*. Diese lässt sich insgesamt 170-mal belegen, und unter den Verwendungsweisen sind zumindest einige,[23] die, wie in den folgenden Beispielen, die abschwächende Funktion eines Diskursmarkers erfüllen.

0006 S1 ich hätte gern mal eine (-) einen rat von ihnen oder ein ihre ansicht gehört.
0007 S2 mhm .
0008 S1 und und zum folgenden problem wir haben uns gestern im familienkreis so darüber gestritten (-) wie weit man äh (-) literatur eh na (-) sagen wir sexueller art (-) wie äh ... oder dieser art.
0009 S2 also aufklärungsbücher **sagen wir mal**.
0010 S1 ja aufklärungsbücher.
0011 S2 is ja weniger literatur als äh wissenschaftliche (-).
0012 S1 ja.
0013 S2 mehr oder weniger wissen.
0014 S1 ... wissenschaftliche (-).
0015 S2 ja.
0016 S1 äh sachen wie weit man die offen im bücherschrank hinstellen sollte also kindern zugänglich (--).

[DS-- E 00005 SE 01 T 01: Hörfunksendung]

0021 S2 ist nicht die heute vollzogene Regierungsbildung einer Kleinen Koalition in Düsseldorf aus SPD und FDP ein gewisses Zugeständnis an diese nun **sagen wir mal** kritische Stimmung in ihrer Partei?

[FR-- E 00017 SE 01 T 01: Fernsehsendung]

0129 S2 Sie haben sich eben wehren müssen da unten, nicht.
0130 S3 Ha ja.
0131 S2 Ja, ja. Ja. Ja, also sie waren aber auch damals schon, das ist ja heut' auch mein äh (PAUSE) das ist vielleicht der Selbsterhaltungstrieb, sie waren so, **sagen wir mal**, ein bißle e profitlich zum Teil, nicht, das ist ja das, was du da vorhin schon angedeutet hast, e ad rem suam attendi, steht in den Büchern drin.

[ZW-- E 00578 SE 01 T 01: Biographische Erzählung]

Wir halten also fest, dass sich nur die Form *sagen wir mal*, nicht aber *ich sag mal* oder *sag ich mal* in den vor 1980 erhobenen AGD-Korpora nachweisen lässt. Angesichts der Menge der durchsuchten Daten und der Tatsache, dass die Form in FOLK (also ab 2006) häufig nachweisbar ist, können wir mit einiger Sicherheit behaupten, dass die Form im von diesen Korpora repräsentierten Sprachgebrauch vor 1980 nicht existierte. Die Korpora decken unterschiedliche Gesprächstypen

[23] Insgesamt scheint nach dem ersten Eindruck aber eine Lesart im Sinne von "zum Beispiel" zu überwiegen, die sich auch in FOLK, dort aber weniger prominent, nachweisen lässt (s.o.).

und eine große regionale Variation ab, allerdings wurden sie allesamt im Gebiet der damaligen Bundesrepublik (und zu kleineren Teilen in Österreich, der Schweiz und Liechtenstein) erhoben und enthalten keine Daten aus dem Gebiet der ehemaligen DDR. Vorsichtigerweise sollten wir unsere Verallgemeinerung also auf das mündliche Deutsch "im Westen" beschränken. Korpora, die den mündlichen Sprachgebrauch in der DDR vor 1980 dokumentieren, gibt es in der DGD noch nicht.[24] Die derzeit einzige Möglichkeit, auch solche Daten einzubeziehen, besteht in einer Abfrage an das in den 1990er Jahren erhobene Berliner Wendekorpus.

4.3.2. Berliner Wendekorpus

Das Berliner Wendekorpus wurde zwischen 1992 und 1995 von Norbert Dittmar im Projekt "Kollektives Gedächtnis – sozialer und sprachlicher Wandel in der Nachwendezeit" an der Freien Universität Berlin erhoben. Es dokumentiert in Form von freien Interviews mit Ost- und West-Berlinern den gesellschaftlichen Umbruch nach dem Fall der Berliner Mauer 1989 als Kollektion individueller und gruppenspezifischer Erfahrungen (Dittmar/Bredel 1999). Für unsere Fragestellung besonders interessant ist, dass das Korpus Äußerungen von ost- und westdeutschen Sprechern in vergleichbaren Gesprächssituationen enthält.

Für die Anfrage ans Berliner Wendekorpus wählen wir dieses in der Korpusliste links aus und führen die oben beschriebenen Schritte aus. Da es sich um ein relativ kleines Korpus (256.924 transkribierte Tokens) handelt, müssen wir die Suche nicht zu restriktiv gestalten und wählen zunächst dieselben Parameter wie bei FOLK (Suche nach Lemma *sagen*, Filtern nach regulärem Ausdruck "(ein)?mal" im beidseitigen Kontext mit Abstand von 3 Tokens). Die verbleibenden 330 Ergebnisse filtern wir noch einmal nach der normalisierten Form *ich* im beidseitigen Kontext mit Abstand von 3 Tokens. Wir erhalten 171 Ergebnisse, von denen – wie ein kurzer Blick auf die zufällig durchmischte KWIC-Konkordanz bereits eindeutig zeigt – ein großer Teil unter die uns interessierenden Verwendungen zu rechnen sind.

	Ereignis	Sprecher		Treffer	
1	BW--_00044	ANTON	uns einklich das was wir ja **ich**	sach	das jetz **ma** so ganz hochgestochn historische aufgabe
2	BW--_00002	DIRK	dir janz ehrlich sicherlich hat man heute	sag	**ick mal** mehr ` möglichkeiten dit jibt sicher viele
3	BW--_00010	GITTA	pf: ja **ick**	sag	**ma** die ägypter hier drüben^ die sind eigentlich
4	BW--_00012	KIRA	wohltuend einfach **ick**	sach	**mal** der is bestümmt sowieso 20 jahre jünger als der
5	BW--_00002	DIRK	ne^ und daß et für uns ooch	sag	**ick mal** n bißel schwerer is einfach zu sagen naja
6	BW--_00043	PIET	unterricht und so weiter stark kontrolliert	sag	**ich mal**.
7	BW--_00019	VERA	her jetzt so grammatikalisch **mal** jesehn wir	sagen	ja **ich** stehe hier oder ich habe dort gestanden^
8	BW--_00028	PETER	die uns **ich**	sag	**mal** ossis damals beklatscht ham wie wir dort über
9	BW--_00002	DIRK	wenn man keene ahnung von nüscht hat	sag	**ick mal** und nich clever jenug is sich so zu

Abbildung 21: Kookkurrenzen von *ich, sagen* und *mal* im Berliner Wendekorpus

[24] Das Korpus "Deutsche Mundarten: DDR", eine variationslinguistische Erhebung mit gut 1600 Aufnahmen, wird im Laufe des Jahres 2015 über die DGD bereitgestellt werden. Es ist allerdings nur zu einem kleinen Teil transkribiert.

Offenbar sind also – im Kontrast zu Westdeutschland vor 1980 – Verwendungen von *ich sag mal* und *sag ich mal* im Berlin der 1990er Jahre keineswegs selten. Die Erklärung, dass die Konstruktion generell auf die Berliner Region beschränkt ist (und deshalb in DS, FR, OS, PF, ZW evtl. nicht auftaucht), haben wir anhand von FOLK bereits ausgeschlossen. Wir überprüfen nun einen Zusammenhang mit der Herkunft der Sprecher aus West- oder Ostdeutschland. Dazu lassen wir uns über METADATEN die Eigenschaft "S: Wohnort (Ortsteil)" einblenden, die für das Wendekorpus den Stadtteil bzw. die Stadtteile dokumentiert, in denen die betreffenden Sprecher zu Hause sind. Ein Sortieren und Auszählen der KWIC ergibt folgende Zahlen:

Stadtteil	Ost/West	Anzahl Belege[25]	Transkribierte Tokens	rel. Häufigkeit
Friedrichshain	Ost-Berlin	56	15278	0,3665%
Treptow	Ost-Berlin	18	8038	0,2239%
Mitte	Ost-Berlin	28	18838	0,1486%
Hellersdorf	Ost-Berlin	30	37053	0,0810%
Marzahn	Ost-Berlin	11	19263	0,0571%
Köpenick	Ost-Berlin	4	8071	0,0496%
Lichtenberg	Ost-Berlin	2	24079	0,0083%
Ost-Berlin gesamt		**149**	**130620**	**0,1141%**
Zehlendorf	West-Berlin	13	9314	0,1396%
Kreuzberg	West-Berlin	1	892	0,1121%
Wilmersdorf	West-Berlin	3	4203	0,0714%
Neukölln	West-Berlin	4	14718	0,0272%
Lichtenrade	West-Berlin	4	18597	0,0215%
Hennigsdorf	West-Berlin	1	8489	0,0118%
Grunewald	West-Berlin	0	3914	0,0000%
Lichterfelde	West-Berlin	0	4305	0,0000%
Mariendorf	West-Berlin	0	6533	0,0000%
Steglitz	West-Berlin	0	7925	0,0000%
Tegel	West-Berlin	0	2615	0,0000%
Spandau	West-Berlin	0	2347	0,0000%
West-Berlin gesamt		**26**	**83852**	**0,0310%**

Tabelle 4: Treffer nach Stadtteilen in Ost- und Westberlin

Auch wenn diese Zahlen nicht hundertprozentig eindeutig sind, so ist die Tendenz doch äußerst deutlich: der weitaus größere Teil der Belege von *ich sag mal*

[25] Hinzu kommen zwölf Belege der Interviewerinnen, für die keine Metadaten dokumentiert wurden. Bei einigen wenigen Sprechern ist mehr als ein Stadtteil als Wohnort dokumentiert. Daraus erklärt sich, dass sich die Belege hier zu 175 aufsummieren, während die KWIC eigentlich nur 171 Belege enthält.

stammt von Sprechern aus dem ehemaligen Ost-Berlin.[26] Dies gilt auch dann noch, wenn wir die absoluten Zahlen mit der Zahl der jeweils transkribierten Tokens in Bezug setzen.

Wir haben hier also einen starken Hinweis darauf vorliegen, dass die Form *ich sag mal* vor der Wendezeit bereits in der DDR (oder mindestens: Ost-Berlin) gebräuchlich war und sich seitdem erst, dann allerdings flächendeckend, im gesamtdeutschen Sprachgebrauch etabliert hat.

4.4. Zusammenfassung

Da in den vorhergehenden Abschnitten großer Wert auf eine detaillierte Darstellung der einzelnen Analyseschritte und ihrer Umsetzung in der DGD gelegt wurde, seien die wichtigsten Ergebnisse der Korpusrecherche nach *mal sagen* hier noch einmal in kürzerer Form zusammengefasst:

- Die Konstruktion lässt sich in der von Auer/Günthner (2003) beschriebenen Form und Funktion problemlos in FOLK nachweisen. Die Häufigkeit der Belege und die Tatsache, dass sie nicht auf einen bestimmten Gesprächs- oder Sprechertyp beschränkt sind, sprechen dafür, das Phänomen als ein häufiges – nicht gelegentliches – zu charakterisieren.
- Die Form der Konstruktion ist nicht vollständig fixiert. Es finden sich Varianten in der Wortstellung, in Modus und Finitheit sowie eine verwandte, zumindest teilweise funktionsgleiche, Konstruktion in der ersten Person Plural.
- Die verschiedenen Formen finden sich in einer Vielzahl syntaktischer Positionen. Es gibt keinen klaren Hinweis darauf, dass sie bevorzugt in syntaktisch peripheren Positionen auftreten.
- Die Konstruktion hat einige typische Begleiter. Es sind dies zum einen die weitestgehend semantisch "leeren" Partikeln *so*, *einfach* und *jetzt*, zum anderen eine (offene) Klasse von Adverbien bzw. Adverbialen, die eine zusätzliche (in gewisser Weise redundante) Motivation für die Verwendung von *mal sagen* liefern.
- Funktional ist allen Verwendungen gemein, dass sie eine Abschwächung des durch *mal sagen* markierten Ausdrucks leisten. Der Grund für die Abschwächung variiert im Einzelfall, hat aber immer mit der Distanzierung des Sprechers von einem Ausdruck zu tun, den er als potentiell problematisch einstuft, z.B. weil er mit negativen Konnotationen oder Tabus verbunden ist, eine über den konkreten Fall hinausgehende Evaluation impliziert oder schlicht unpräzise ist. Durch vorangestelltes oder nachgeschobenes *mal sagen* signalisiert der Sprecher seinem Gesprächspartner, dass er die markierte Formulierung als vorläufig und damit gegebenenfalls später noch verhandelbar verstanden wissen möchte.

[26] Dies mag auch für 12 der 13 Belege aus Zehlendorf gelten, denn für die betreffende Sprecherin mit dem Pseudonym Gitta ist außer Zehlendorf auch Treptow als Wohnort dokumentiert. Nimmt man nur Belege, bei denen die Zuordnung nach Ost- oder Westberlin eindeutig ist, wird die Tendenz noch deutlicher.

- Die Tatsache, dass sich die Form in der ersten Person Singular anhand einer großen Datenbasis westdeutscher Daten vor 1980 gar nicht nachweisen lässt, sowie die deutlichen Unterschiede in der Frequenz zwischen Ost- und West-Berliner Sprechern aus den frühen 1990er Jahren lassen vermuten, dass die Formen *ich sag mal* und *sag ich mal* ihren Ursprung im Sprachgebrauch der DDR haben und sich erst seit der Wende auch im westdeutschen Sprachgebrauch flächendeckend etabliert haben.

5. Fazit und Ausblick

Wie dieser Beitrag zu zeigen versucht hat, lassen sich anhand eines großen Gesprächskorpus und einer Datenbank, deren Funktionalität gesprächsanalytische Arbeitsweisen unterstützt, Erkenntnisse gewinnen, die die Ergebnisse "klassischer" gesprächsanalytischer Analysen zu ergänzen, teilweise auch zu präzisieren oder korrigieren vermögen. Insbesondere erlaubt der Rückgriff auf eine große, einheitlich erschlossene und systematisch dokumentierte Datenbasis wie FOLK zu beurteilen, wie typisch eine Einzelfallanalyse für ein untersuchtes Phänomen ist, und wie weit und mit welchen Einschränkungen oder Ausdifferenzierungen sich ihre Ergebnisse generalisieren lassen. Ein ausreichender Umfang und eine große Variationsbreite der Datenbasis sind dafür notwendige, aber nicht hinreichende Voraussetzungen. Das Potential dieser Datenmengen wird erst durch eine geeignete texttechnologische Erschließung und durch die automatisierten Recherchefunktionalitäten einer Datenbank auch praktisch nutzbar. Wenn sich diese automatisierte Analyse unmittelbar und ohne praktische Hürden wieder mit einer detaillierten ("qualitativen") Studie der Daten rückkoppeln lässt – so wie es die DGD über den Rückgriff auf Transkript und Audio aus KWIC-Konkordanzen heraus zu ermöglichen versucht – müssen etablierte gesprächsanalytische Arbeitsweisen dafür nicht aufgegeben werden, sondern können sich mit korpuslinguistischen Herangehensweisen an Gesprächsdaten sinnvoll und gewinnbringend kombinieren lassen.

Die in diesem Beitrag dargestellte Analyse ist hierfür nur ein illustrierendes Beispiel, das keineswegs den Anspruch erhebt, das Phänomen selbst oder das Potential korpus- und datenbankgestützter Gesprächsanalysen erschöpfend beschrieben zu haben.

Beispielsweise könnt die Analyse von *mal sagen* dahingehend erweitert werden, dass auch noch verwandte Konstruktionen mit anderen verba dicendi (*ich behaupte mal*, *ich formulier's mal so* – beide in FOLK belegt) oder Mentalverben (*ich denke mal* – ebenfalls in FOLK belegt) in einer analogen Herangehensweise untersucht werden, oder es könnte anhand der Audiodaten analysiert werden, welche Merkmale und welche Variation sich auf der prosodischen Ebene für diese Formen feststellen lassen.

Weiterhin hat die Analyse auch deutlich gemacht, dass anhand der derzeit in der DGD verfügbaren Datenbasis bei Weitem nicht alle Fragen beantwortet werden können, die sich bei einer korpusgesteuerten Analyse des Phänomens aufdrängen. Interessant wären in diesem Zusammenhang etwa der Einbezug weiterer Gesprächstypen (z.B. die von den Studierenden genannten Fußballer-Interviews

oder Talkshows) oder die Konsultation von mündlichen Daten aus der DDR vor 1980.[27]

Schließlich zeigt die Analyse nicht nur die Möglichkeiten, sondern auch einige praktische Grenzen der datenbankgestützen Analyse auf. So könnten etwa einige der quantitativen Auswertungen, die hier weitestgehend händisch erledigt wurden, sicherlich sehr effizient durch automatische Methoden unterstützt werden, und es ist bereits absehbar, dass manche Beschränkungen der Datenbank – wie etwa die Begrenzung auf 10.000 Suchergebnisse – sich zukünftig als hinderlich erweisen können, wenn die Datenbasis wie geplant weiter wächst.

Da sich sowohl FOLK als auch DGD noch mitten im Aufbau befinden, werden zumindest einige dieser zusätzlichen Möglichkeiten in Zukunft (etwa mit der Integration des DDR-Korpus in die Datenbestände, der kontinuierlichen Erweiterung von FOLK um zusätzliche Gesprächstypen und dem Ausbau der Datenbankfunktionalität) umsetzen lassen.

6. Literatur

Anthony, Laurence (2013): A critical look at software tools in corpus linguistics. In: Linguistic Research 30, 2, 141-161.

Auer, Peter / Günthner, Susanne (2003): Die Entstehung von Diskursmarkern im Deutschen - ein Fall von Grammatikalisierung? Interaction and Linguistic Structures (InList) 38.
http://www.inlist.uni-bayreuth.de/

Biber, Douglas (2004): Representativeness in corpus design. In G. Sampson / D. McCarthy (Eds.), Corpus linguistics: Readings in a widening perspective, London: Continuum,174-97.

Brinckmann, Caren / Kleiner, Stefan / Knöbl, Ralf / Berend, Nina (2008): German Today: an areally extensive corpus of spoken Standard German. In: Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesch, Marokko.

Bührig, Kristin / Kliche, Ortrun / Meyer, Bernd / Pawlack, Birte (2012): The corpus "Interpreting in Hospitals": Possible applications for research and communication training. In: Schmidt, Thomas / Wörner, Kai (Hg.): Multilingual corpora and multilingual corpus analysis. Amsterdam: Benjamins, 305-318.

Deppermann, Arnulf / Hartung, Martin (2011): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des "Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder, E. / Müller, M. / Vogel, F. (Hrsg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin/New York: de Gruyter, 414-450.

Deppermann, Arnulf / Schmidt, Thomas (2014): Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik - Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes

[27] Interessant könnte darüber hinaus ein Einbeziehen größerer Mengen von (in AGD, FOLK und DGD unterrepräsentierten) Daten aus der Schweiz und aus Österreich sein. Eventuell lassen sich dort Unterschiede in der Verbreitung von *ich sag mal* feststellen, weil der Sprachgebrauch der DDR dort nicht in gleicher Weise Eingang in die öffentliche Wahrnehmung gefunden hat wie in Deutschland.

Deutsch (DGD2). In: Mitteilungen des deutschen Germanistenverbandes 1 (2014), [Domke, Christine / Gansel, Christa (Hg.): Korpora in der Linguistik - Perspektiven und Positionen zu Daten und Datenerhebung], 4-17.

Dittmar, Norbert / Bredel, Ursula (1999): Die Sprachmauer. Die Verarbeitung der Wende und ihre Folgen in Gesprächen mit Ost- und WestberlinerInnen. Berlin: Weidler Buchverlag.

Fandrych, Christian / Meißner, Cordula / Slavcheva, Adriana (2012): The GeWiss Corpus: Comparing Spoken Academic German, English and Polish. In: Schmidt, Thomas / Wörner, Kai (Hg.): Multilingual corpora and multilingual corpus analysis. Amsterdam: Benjamins, 319-337.

Fiehler, Reinhard / Wagener, Peter (2005): Die Datenbank Gesprochenes Deutsch (DGD) - Sammlung, Dokumentation, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. Gesprächsforschung (6), 136-147.

Gasch, Joachim / Brinckmann, Caren / Dickgießer, Sylvia (2008): memasysco: XML schema based metadata management system for speech corpora. In: Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesch, Marokko.

ten Have, Paul (1990): Methodological issues in conversation analysis. In: Bulletin de Méthodologie Sociologique 27, 23-51.

Hunston, Susan (2002): Corpora in applied linguistics. Cambridge: Cambridge University Press.

Kallmeyer, Werner (2007): Möglichkeiten der maschinellen Unterstützung bei der Arbeit mit Interaktionskorpora. In: Kallmeyer, Werner / Zifonun, Gisela (Hrsg.): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Berlin/New York: de Gruyter, 203-234.

Kupietz, Marc / Schmidt, Thomas (erscheint): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. Beiträge zur Jahrestagung 2014 des Instituts für Deutsche Sprache.

Lemnitzer, Lother / Zinsmeister, Heike (2006): Korpuslinguistik – eine Einführung. Tübingen: Gunter Narr.

Schmidt, Thomas (2003): Korpus "Skandinavische Semikommunikation" - ein mehrsprachiges Diskurskorpus auf XML-Basis. In: Sprachtechnologie für die multilinguale Kommunikation - Textproduktion, Recherche, Übersetzung, Lokalisierung. Beiträge der GLDV-Frühjahrstagung 2003 an der Hochschule Anhalt (FH) in Köthen, 421-427.

Schmidt, Thomas (2005): Datenarchive für die Gesprächsforschung. Perspektiven, Probleme und Lösungsansätze. In: Gesprächsforschung (6), 103-126.

Schmidt, Thomas / Schütte, Wilfried (2010): FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction. In: Nicoletta Calzolari et al. (eds.): Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 10), may 19-21, 2010, Valletta, Malta. European Language Resources Association (ELRA), 2091-2096.

Simpson-Vlach, Rita C. / Leicher, Sheryl (2006): The MICASE Handbook: A Resource for Users of the Michigan Corpus of Academic Spoken English. University of Michigan Press.

Sinclair, John (2004): Trust the text: Language, corpus and discourse. London: Routledge.

Stift, Ulf-Michael / Schmidt, Thomas (2014): Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In: Institut für Deutsche Sprache (Hrsg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Redaktion: Melanie Steine, Franz Josef Berens. Mannheim: Institut für Deutsche Sprache, 360-375.

Weber, Peter (2014): Verkaufsgespräche führen lernen in der Schule. Eine linguistische Untersuchung. Mannheim: Verlag für Gesprächsforschung. http://www.verlag-gespraechsforschung.de/2014/peterweber.html

Westpfahl, Swantje / Schmidt, Thomas (2013): POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: Journal for Language Technology and Computational Linguistics 1, 139-156.

Wiese, Heike / Freywald, Ulrike / Schalowski, Sören / Mayr, Katharina (2012). Das KiezDeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. Deutsche Sprache 2, 97-123.

Dr. Thomas Schmidt
Programmbereich Mündliche Korpora
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim

thomas.schmidt@ids-mannheim.de

Veröffentlicht am 9.3.2015